



A University of Sussex PhD thesis

Available online via Sussex Research Online:

<http://sro.sussex.ac.uk/>

This thesis is protected by copyright which belongs to the author.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Please visit Sussex Research Online for more information and further details

**Anomaly handling: Strategic flexibility in a complex
problem-solving environment**

Robert Leadley

BSc(Hons), MSc

Schools of Psychology

University of Sussex

September, 2016

This thesis is submitted in part fulfilment of the requirements for the degree of Doctor of

Philosophy

Declaration

I declare that this thesis is my own work and has not been submitted in substantially the same form for the award of a higher degree at this institution or elsewhere.

Robert Leadley

Abstract

This thesis aimed to develop a paradigm for the study of anomaly handling and to investigate the factors that influence success in detecting and classifying anomalies. A simulated anomaly-handling environment was created to mimic an intelligence analysis task in a security setting. A series of experiments was designed to test hypotheses concerning sources of difficulty in detecting potential anomalies and making decisions about appropriate classifications of potentially anomalous events. Results across all experiments showed that complex problems, representing anomalies, were more difficult to solve than simple problems, and that this poor performance was consistent with the use of suboptimal strategies based on recognition of perceptual characteristics rather than inferences drawn from available data. Performance on complex problems was reduced still further when participants were exposed to trials that established a mental set. However, performance was improved when participants were given feedback on the correctness of their responses to each trial, which eliminated the negative effects of exposure to mental set. Another factor that impacted on successful decision-making was the cost of making errors. When participants were faced with a penalty for making incorrect decisions, solution rates improved compared with when performance was not related to reward. This has consequences for anomaly handling industries where the consequence of failure is often high. Unexpectedly, a number of the results indicated that there are situations where mental set may confer a benefit to decision making in a task of anomaly categorisation. Given the dominance of recognition-based strategies, it appears that mental set can refine the detection of perceptually relevant patterns, which can signal sudden changes in pattern that can lead to a switch from recognition-based to inferential task solution strategies. Overall, the merits for the use of simulated environments in critical decision making areas are discussed, and the contributory factors towards successful anomaly handling are analysed.

Acknowledgements

I would like to thank my supervisor, Professor Thomas Ormerod for giving me the chance to test myself in this endeavour. Your insight, knowledge, wisdom and patience seemingly knows no bounds, and I appreciate the time and energy you invested in me.

I would also like to thank my wife, Dr Sarah Garner, for her unending emotional support and patience through the difficult stages of this work.

I would also like to thank Dr Alex Sandham and DSTL; without their generous funding and specialist topic knowledge this project would not have a home.

Contents

1	Introduction.....	9
1.1	Overview and scope.....	9
1.2	Decision-making.....	14
1.2.1	What goes into a decision	14
1.2.2	Fallibility in human decision-making	25
1.3	From laboratory to field – Human involvement in anomaly handling	30
1.3.1	Anomalies in naturalistic environments.....	30
2	Developing a simulation for exploring anomaly handling	39
2.1	An introduction to the experimental paradigm	39
2.1.1	Overview and goal of the simulation	39
2.1.2	Information available within the simulation	40
2.2	Technical rules overview	43
2.2.1	Environmental events.....	43
2.2.2	Conflict events	46
2.2.3	Adding complexity.....	48
2.2.4	Inferences	50
2.3	The use of simulations in psychological research.....	54
3	Experiment 1: Initial performance evaluation	57
3.1	Introduction.....	57
3.2	Method	61
3.2.1	Participants.....	61
3.2.2	Design	61
3.2.3	Materials	62
3.2.4	Procedure	64

	3.2.5 Coding Strategy	65
	3.3 Results.....	66
	3.3.1 Analytical approach	66
	3.3.2 Solution rates	67
	3.3.3 Solution Times	68
	3.3.4 Strategy use.....	69
	3.3.5 Strategy and Response times	72
	3.4 Discussion.....	73
4	Experiment 2 – Feedback.....	80
	4.1 Introduction.....	80
	4.2 Method	81
	4.2.1 Participants.....	81
	4.2.2 Design	82
	4.2.3 Materials	82
	4.2.4 Procedure	82
	4.3 Results.....	83
	4.3.1 Solution rates	83
	4.3.2 Solution times	84
	4.3.3 Learning over trials	85
	4.4 Discussion.....	86
5	Experiment 3 - Cost	90
	5.1 Introduction.....	90
	5.2 Method	91
	5.2.1 Participants.....	91
	5.2.2 Design	91

	5.2.3	Materials	92
	5.2.4	Procedure	92
	5.3	Results.....	93
	5.3.1	Solution rates	93
	5.3.2	Solution times	95
	5.4	Discussion	97
6		Experiment 4: Manipulating mental set.....	100
	6.1	Introduction.....	100
	6.3	Method	101
	6.3.1	Participants.....	101
	6.3.2	Design	102
	6.3.3	Materials	102
	6.3.4	Procedure	103
	6.4	Results.....	103
	6.4.1	Solution rates	103
	6.4.2	Solution times	105
	6.5	Discussion	106
7		Experiment 5 –Protocol analysis of strategies.....	110
	7.1	Introduction.....	110
	7.2	Method	112
	7.2.1	Participants.....	112
	7.2.2	Design	112
	7.2.3	Materials	112
	7.2.4	Procedure	112
	7.2.5	Verbal protocols and cognitive operations	113

7.3	Results.....	115
7.3.1	Cognitive Operations	115
7.3.2	Analysis of failed decisions	116
7.3.3	Creating a strategic model to represent decision making	120
7.4	Discussion	121
8	General Discussion	124
8.1	Aims and Overview	124
8.2	Establishing the anomaly handling paradigm	130
8.3	Boundaries of successful anomaly handling.....	133
8.4	Applications	139
8.5	Future research.....	140
8.6	Conclusion	141
9	References.....	142
	Appendix 1. Screenshots of the online task instructions	149
	Appendix 2. Trial order for Experiment 4.....	152

1 Introduction

1.1 Overview and scope

In July 2008, the United States of America conducted a drone strike in Nangarhar province Afghanistan, after spotting a large convoy of individuals believed to belong to terrorist organisations. However, after the strike hit the target, these individuals were later found to have no links to such organisations and were in fact travelling to a local wedding. This attack killed 47 civilians including the bride, with most victims being women and children (Sturke, 2008). In another incident in November 2008, after fighting insurgents near a town in Kandahar province, Afghanistan, a US drone strike targeted a nearby building containing multiple occupants believed to be those same insurgents. The drone strike hit the compound, but the building was being used by women and children to celebrate a local wedding. The strike killed 37 civilians (Wafa & McDonald, 2008).

Drone strikes offer a fast, precise response to threats, minimising the risk to ground troops, so it is no wonder they are well utilised. However, they are only as effective as the decisions to employ them. In both the above examples, the whole picture was incomplete, and it led to a large loss of civilian life. It may have been that those strikes were ‘signature strikes’ – where the strike decision is based on a pattern of suspicious behaviour, rather than on gathered intelligence that indicates a specific target (Ackerman, 2016).

In signature strikes, a profile of suspicious behaviour is used in the decision to strike. But how do you tell the difference between a wedding convoy moving between villages and an armed convoy on its way to recruit more insurgents? What does a house full of insurgents look like compared to a house full of wedding revellers? It is possible that while insurgent activity is common and under constant scrutiny, drone operators and intelligence officials have learnt what insurgency ‘looks like’, consisting of patterns of human behaviour that usually correlate with insurgent activity. However, there may be less frequent events, such as

a wedding, that does not happen on a day-to-day basis and where the conditions appear very similar to that of insurgency. Large groups of people travelling large distances and congregating in one place may be suspicious to the military in the majority of instances, but far less frequently this pattern may also match something of an innocent nature, such as a wedding party. In both instances, we can observe an unusual change in the environment, maybe a number of vehicles travelling in convoy, but how do we make a decision about what that represents? And is the military nature of drones and their familiar use to combat insurgents going to impact on the way a drone operator may interpret that situation?

This is the essence of anomaly handling – when a change is noticed, how is it interpreted when there are multiple possibilities, and how does this change when some possibilities are far more probable than others? While rare innocent events may appear like more common events of a suspicious nature, it is perhaps no surprise that mistakes are sometimes made. When combat policy such as signature strikes rely on human interpretation, there is always the possibility that the human gets it wrong. It is perhaps in this failure to distinguish between two similar looking events that has created an environment where drone strikes are able to hit weddings on multiple occasions. News sources found that US drones had hit eight wedding parties between 2001-2013 (Englehardt, 2013), and while these fortunately are rare events, the repeated nature of such tragedies may imply a procedural reason for why this is occurring. If we were to better our understanding of how anomalies such as these are handled and processed by decision-makers, it may be possible to explore what options exist to reduce the likelihood of errors. As of mid-2016, signature strikes are still utilised by the US military drone program (Ackerman, 2016).

In a wider context, anomalies are events that are unexpected and do not fall within normal parameters. It is important to understand the ability to interpret and comprehend anomalies because they are often indicative that a system is deviating from its normal state

(e.g., irregular sensor data from a power plant may be symptomatic of an undetected problem, or a sudden large convoy of fast moving vehicles may indicate an impending insurgent attack). By being able to interpret anomalies successfully, much of the potential damage posed by these threats may be prevented or mitigated. Many industries use analysts to scrutinise data for anomalies in an attempt to identify problems in such systems (e.g., in the control of process plants) or to counter malicious activity (e.g., fraud investigators, military intelligence).

Automated anomaly detection systems have improved security in many industries. Even though automated systems are typically closely supervised by human decision-makers, there is little psychological literature on humans' capacity to interpret these anomalies. This thesis aims to explore the decision-making processes of anomaly handlers, and to identify key psychological mechanisms that allow successful interpretation of anomalies to occur.

To explore the concept of anomaly handling, a simple computerised problem-solving task was developed. The task involves participants finding and interpreting a pattern of anomalies within a simulated security environment. This environment simulates a domain in which our concept of anomaly handling is used regularly and has been discussed already, that of military intelligence. The intelligence context is one where data will be analysed with the intent of discovering patterns of behaviour that might be associated with malicious intent, while facing the same potential problems of discerning a threat from a non-threat when the two look very similar. The research follows on from qualitative research already undertaken to examine effectiveness of problem solvers in intelligence analysis (e.g., Mumaw, Roth, Vincente, & Burns, 2000; Morely, Ball, & Ormerod, 2006). The objective of this research was to devise and use a method that can assess anomaly-handling abilities quantitatively in a laboratory setting. The process of creating the simulated environments, demonstrating the requirements of the task, and how the design of the task meets these requirements is

addressed in Chapter 2.

One of the major themes reviewed in this thesis is decision-making, and most importantly, understanding why it goes wrong. A literature review of decision-making and relevant cognitive biases can be found in the current chapter. Some key areas are outlined below to develop the perspective from which this work is focussed.

Kahneman and Tversky (1973a; 1973b) and Tversky and Kahneman (1974) developed a series of theories to explain why human decision making that involves uncertainty and concerns risk is often subject to a number of cognitive biases that arise through the application of heuristics. Heuristics are mental rules-of-thumb, or shortcuts, formed by experience and are beneficial by reducing cognitive load. While heuristics have been shown to have advantages for rapid processing of routine decisions, they may create problems for anomaly interpretation. If heuristics are cognitive shortcuts based on experience, then the low frequency of some anomalous events means that decisions based upon heuristics formed from common experience may impede the identification and interpretation of an anomaly. For example, if every vehicle convoy a drone operator has seen so far results in an insurgent attack, then the operator may associate convoys with insurgent attacks. However, there may be other causes (such a wedding) that creates similar signals. As the operator is exposed to this on a much less frequent basis then it is understandable how a misattribution could occur. Literature on how error can be introduced into rational decision-making is discussed in Section 1.2.2.

Alternatively, anomaly handlers may develop new heuristics to combat the uncertainty of environments in which they work, which may improve performance. In the drone example, after a wedding has been hit, those decision-makers will be aware that the behaviour used to determine the strike criteria was inaccurate, and may develop further criteria to be sure their next strike doesn't make the same mistake. After viewing enough

satellite imagery of typical wedding behaviour, they will be able to develop a more refined set of behaviour that could distinguish between weddings and insurgents. This view is supported by the literature on expertise, where research in the field of Naturalistic Decision Making (NDM) suggests that experts in their domain develop decision-making strategies exclusive to that domain (Klein, Orasanu, Calderwood, & Zsombok, 1993). A review of the literature on expertise and its boundaries can be found in Section 1.3.

Although these global aspects of anomaly handling are important, we are interested in specific elements that fit with the task paradigm chosen for our research. The task allows investigation of the contribution that factors such as mental set, feedback and cost of error have on anomaly handling. By manipulating these factors, their effects on performance and strategy choice can be examined to determine how critical those factors are for successful anomaly handling.

Mental set, traditionally known as the *Einstellung effect*, was first described by Luchins (1942). A mental set is associated with a rigid approach to problem-solving, often with fixation on previously used strategies that have yielded consistent positive results. Mental set is important in anomaly handling because by their very nature, anomalies are different to what has come before them. Individuals whose primary job is to carry out similar duties in their day-to-day working environments may become susceptible to creating a mental set based around their workplace actions, and thus become less sensitive to the presence of anomalies and their consequences. For example, if the signature strike procedure is used by drone commanders to target a large group of individuals in vehicles, and is successful on the first 10 occasions, this may increase the likelihood of assessing the next large group to be valid targets when it is in fact yet another wedding party. The issue of mental set is further addressed in section 1.2.2.

It is also important to examine the role of thinking styles. Human thinking has been

described using a dual-process model, identifying heuristic and analytic types of thinking (Evans, 1989; Kahneman & Frederick, 2002; Kahneman, 2003). Heuristic thinking is more accessible, and expressed through quick, routine decision-making, whereas analytic thinking is slower, and utilises more cognitive resources. It could be that switching between heuristic and analytic modes of thinking may lead to a reduction in errors. Factors such as introduction of a cost to incorrect decisions or providing feedback to participants may encourage participants to use analytic strategies that lead to more efficient strategy selection and thereby enhance anomaly handling performance, and these are explored in the experiments reported in this thesis.

The aim of the thesis research was to develop a novel method to investigate the nature of anomaly handling, capable of identifying the psychological mechanisms used when undertaking anomaly handling tasks. The research provides quantitative data that explores novice performance and strategy selection in anomaly handling.

1.2 Decision-making

1.2.1 What goes into a decision

Possession of information is critical for a decision. Without information a decision is little more than a guess. So how do we determine what information is necessary before a successful decision can be made? Dawes and Corrigan (1974) suggested that the key to successful decisions is to determine cues to necessary task information. Cues are features in an environment from which relevant information can be extracted dependent on the perceivers' ability to interpret them. For example, if a drone operator was to identify hostile groups, a cue that there is a threat may include the number of people in an area, the behaviour of those people, whether they are carrying weapons, or the mode of travel. These are all cues that an experienced operator can use to form an assessment of the threat on the ground. These are cues as they only imply the nature of the event, in this case, threat. So a cue is an aspect

of the environment from which useful information can be obtained, an informational object. In the case of decision-making, decisions are often based upon the information extracted from these cues.

Dawes and Corrigan (1974) suggest that the key to decision-making is an individuals' ability to perceive their environment and to use the relevant cues to extract the most pertinent information related to a specific decision. To return to our drone example, if an operator has to decide that a situation is dangerous and that a strike should go ahead, then relevant cues would not only include the behaviour of targets, and presence of weapons, but also extend to whether nearby friendly troops are out of range, or whether there are a large number of civilians in the area. The tasks of identifying hostiles and conducting a strike utilise information available from the environment, but each may require different information. Cues are important because they impart information to us that can better our understanding of a situation or problem. The environment is filled with cues; learning to discriminate what is an important cue from an irrelevant cue is critical in decision-making and anomaly handling. This is the basis for using a simulation in our research with multiple information streams which will act as cues. By creating multiple information streams it is possible to manipulate what information is relevant or irrelevant to solving a problem. This will allow those with the ability to distinguish between the relevance of cues to perform better than those that do not.

How people identify the relationship between cues and their outcomes is the starting point for this thesis. The real world is a noisy, messy place with many simultaneous cues, so how do individuals' learn which cues are valuable? How do people search their environments for the cues that they can extract the information they are looking for? Laboratory studies conducted on information search often examined the concept of 'cue validity', the extent to which a cue can be predictive of a particular outcome. This is commonly explored using a Multi-Cue Probability Learning (MCPL) paradigm (e.g., Friedman & Massaro, 1998), where

two or more cues are presented, usually perceptual, consisting of shapes, colours, etc. Participants must then predict an outcome, such as movement, or appearance of consequent stimuli. Participants are able to predict the outcome from the cues because each has a specific weighting which is predictive of an outcome, and participants learn these weightings over multiple trials. Friedman and Massaro used a medical diagnosis task in which participants had to predict one of two diseases (the outcome) from a patient's blood pressure and temperature (the cues). They found that, even though the rate of occurrence of each disease was different, participants were able to extract relevant information from the cues and became efficient at diagnosis.

Klayman (1988) found that one of the most important functions of a decision-maker is the natural acquisition of valid cues from the environment. He presented participants with a geometric-spatial task where participants had to estimate the movement of shapes based upon the shapes' physical characteristics, but participants were not explicitly told which features of the shapes were cues to each movement. Not only were participants adept at identifying which cues were relevant, but also when they could test their own hypotheses regarding cue relationships they outperformed those who were merely able to observe the cues. This suggests that individuals can learn causal relationships between abstract stimuli under laboratory conditions. This research is imperative to this thesis as it shows that novice participants are able to understand the role of specific cues in an informational environment and perform tasks based on the relationship of these cues. This ability will be required to solve the problems that appear in our experimental paradigm and that previous research has found novice participants adept at doing so will not justify a poor performance on the task, which means poor performance can be associated with less ability in anomaly handling.

Similarly, Lagnado and Sloman (2004) have shown that intervention, the ability of the participant to have direct input, provides greater basis for learning than just observation in

causal reasoning. They produced a task similar to MCPL, where multiple cues lead to an outcome event, with cues part of either a causal chain, where one event follows another, followed by a third in sequential order ($A \rightarrow B \rightarrow C$) or a common cause for an event where there are two separate events that can independently cause it ($A \rightarrow C \leftarrow B$). Participants either observed multiple presentations of scenarios, or were able to intervene with the task and set their own terms for A and B. They found that intervention (directly manipulating the conditions that lead to the outcome) greatly assisted learning the causal relationship between cues and outcome.

Although intervening with a situation can help individuals to learn cue validity over a number of trials, the scarcity of events in anomaly handling domains means that this kind of learning is not always possible. There are industries where discrete one-off observations provide the only available information. How an individual can learn what cues are predictively valid in a potentially high-risk environment where anomalous events are rare holds the key to how the ability to understand that environment can be improved. In the example scenario used in this thesis, an intelligence officer in a security setting has access to imagery of a remote country via satellite, where satellite images depict resources being distributed towards various population centres. There is a relationship between the distribution of resources and the security situation on the ground. A large build-up of resources in one area could indicate that local rebels are stockpiling munitions, food and medicine in preparation of an impending attack. Alternatively, a build-up of resources may be due to adverse weather events that make some areas unreachable necessitating stock-piling at regions that are accessible. In situations like this, the observer does not have multiple trials to learn relationships prior to showing accurate performance levels, as any error could be costly to security. The observer relies on knowledge and expertise to analyse cues of the environment in making their decision. This scenario is the basis of the simulation used

throughout this thesis to test and understand anomaly handling and is described in detail in Chapter 2. The simulation was designed to represent a mirror of the threat (e.g., insurgents) and no-threat (e.g., wedding party) situations described in the start of this thesis, where a harmful and a peaceful outcome might look similar on initial presentation. For the nature of this task, rather than committing to aggressive action such as a drone strike, the paradigm has been constructed around a scenario of observation, and does not imply the loss of civilian life in case of errors.

Studies that examine information search strategies have focussed on the order in which information is gained, as a measure of the priority that participants place on that item of information. For example, Payne (1976) used the apartment task in which options (different apartments) had features (rent, location, etc.) and participants were told to search this information for their preferred apartments. He found that, while there was variation in search orders between individuals, there was consistency within individuals across searches when comparing between two alternative choices. He also found that search order changed with task complexity. When multiple options were introduced, participants no longer prioritised information to choose between competing options, but instead prioritized information to eliminate weaker options. The apartment task taps into differences in subjective preferences of participants who value different qualities in a home, which explains the lack of consistency between their searches. The security scenario described above has no personal preferences, instead providing a task with well-defined criteria and an optimal solution. Thus, the current research examined the nature of information search in a setting that allowed assessment of inter-individual information search strategies to explore differences between strategic approaches independent of personal preferences. The paradigm included multiple information streams to enable participants to make decisions whilst having all available information that would allow a correct solution to be discovered.

Fasolo, McClelland and Lange (2005) examined consumer choice and information search when purchasing a new item. Similar to the apartment task, participants were presented with a list of cameras, each with numerous features such as cost and optical zoom. However, they used a recommendation system whereby the participants selected the option that best fitted the description given to them, thereby eliminating subjective preferences present in Fasolo et al.'s Apartment task. They found that when an option had conflicting features (i.e., it fit the description in one attribute but not another), search became more difficult, and participants reported less confidence in their choices. Participants' information search strategies focused on features that were considered most important, and included searches for alternative choices that may better match their recommendation. If no option satisfied these criteria, participants' either searched further for more information or attributed weightings of importance to each feature. This is important to the task developed for this thesis as anomalies are situations that will possess conflicting information. Therefore information search will be more difficult and participants may attribute a ranking of importance to the cues available in the task paradigm, this will be important later when discussing the development of heuristic strategies.

Bettman, Johnson, Luce and Payne (1993) outline a trade-off between accuracy and effort in information search, in which individuals either process all information regarding options to make an accurate choice, expending more time, effort and mental resources while doing so, or they focus on highly weighted attributes and identify an acceptable option with lower effort. They found that when faced with a conflicted choice, participants were more likely to pay attention to all of the information in an attempt to process it accurately. When no conflict existed, people used less information and more readily accepted an option as acceptable. Thus, participants appeared willing to expend effort if it increased the chance of making a correct decision, but use heuristics when searching among options for more simple

decisions. Extra effort was only expended if participants noticed a conflict in the information supporting a decision. In some real-life situations, conflicting information may exist but, perhaps due to domain unfamiliarity or complexity, the observer may not be aware of it. If this is the case, then we might expect the individual to utilize heuristics and make suboptimal decisions.

How does this notion of effort and accuracy trade-offs affect anomaly handling? Anomalies are indicated by unexpected cues that indicate a problem. Experience may teach individuals that investment in a particular cue of low informative value is not worth the *effort* of attending to it, and so when this cue changes to become diagnostically relevant it already has a preset low weighting for how useful it is. If individuals do not recognize there is a conflict in the available information, due to the presence of an anomaly, then they will be more likely to process it as a routine problem, and not expend additional cognitive resources to complete a search of all available information, thus ignoring potentially important information that a problem exists.

Fiske and Taylor (1991) found that when overwhelmed with information or when facing a topic that is not fully understood, individuals will seek less information because they are unsure what is diagnostically relevant, and therefore will not wish to expend mental resources for uncertain gain. If domain unfamiliarity reduces the amount of information that is attended to, then it could be expected that familiarity with a domain will allow for high amounts of information to be attended to. Evidence in support of this was found by Phelps and Shanteau (1978). They found that experts are capable of incorporating large amounts of information into their decisions. They studied expert livestock judges in agricultural contests. Their experts were given descriptions of livestock that varied on 11 criteria and were asked to judge the quality of an animal based upon a given description. They found that experts referred to all 11 criteria in their judgements. Phelps and Shanteau also indicated that an

expert can access this information in a schematic fashion. This was evidenced in a second study, where the same livestock judges judged animals based upon photographs, they found experts used a maximum of three characteristics to base their decisions on. They suggested that experts are able to extrapolate multiple features of an animal from a single visible characteristic, suggesting use of a formal knowledge structure. Based on the findings of both Fiske and Taylor (1991) and Phelps and Shanteau (1978), if both experts and non-experts conserve mental energy and have the same motivation to succeed, then a possible reason the expert acquires more information is because the expert's knowledge structure is set up in a way that makes retrieval less taxing.

Experts are not experts because they can find more information, but they find more information because they are expert. A review by Shanteau (1992) found that expert and non-expert populations often utilize the same limited amount of information, yet experts still possess higher fidelity in their judgments and decisions. It appears that experts do not need a greater capacity for information to achieve better decisions, but their success is based instead on the internal weighting of individual cues and interpretation of the cues themselves. To explore this notion in anomaly handling, we developed a simulation that used few cues, in which success was measured by understanding of how cues interact rather than on capacity of memory. This enabled an assessment of the relative importance of quality versus quantity of information used in anomaly handling. This also has the added benefit of not isolating our simulation for use in expert communities only. Following Fiske and Taylor (1991), by using a minimal number of cues, novices will still be able to attend to all information. And given Shanteau's (1992) finding that better decisions can be made with a smaller number of cues, this would still make performance on our task one of better understanding rather than domain familiarity.

Information that has been attended to and acquired can be used to inform a decision.

But in what ways can this information be used? Dawes and Corrigan (1974) suggested it could be as simple as adding all of the diagnostic cues together to come to a conclusion. Studies by Dawes, Faust and Meehl (1989) and Grove and Meehl (1996) have shown that, by using a weighted additive function, whereby each piece of information is assigned a value as to how informative it is and then these are added together, can outperform the judgements of experts. Sarbin (1943) conducted a study that compared school counsellors' predictions of students' academic success against a simple equation. The counsellors had a wealth of experience in managing students, an 8-page document detailing the specifics of each student, access to all previous test scores, results from personality measures, and even an interview with the student before making their prediction. Sarbin found that a two-item equation, using just two of the test scores that the counsellors also had access to created equally powerful predictions of the students' future academic grades. A meta-analysis by Grove, Zald, Lebow, Snitz and Nelson (2000) examined 136 experiments that compared simple statistical models against the judgements of experts in the fields of medicine, mental health and education and found significantly better performances by using additive statistical approaches. But why would an expert perform worse than a simple equation? And why would we continue to require experts when computational alternatives provide more reliable results?

Firstly, experts are human and as such, are affected by things computers are not. The argument has been made that computers will arrive at the same conclusion if the data is similar, whereas human experts have the ability to vary in their performance between cases (Newell, Lagnado, & Shanks, 2007). The variability of experts has been studied recently in a security context in a study by Miller, Appleby, Garibaldi, and Aickelin (2013). They demonstrated that a group of experts had different opinions about the prevalence of particular cyber threats, the ease with which cyber-attacks could be executed, and the relative cost of successful cyber-attacks. This research showed that there is a great deal of inter-expert and

intra-expert variability. Inter-expert variability may be explained by some people being ‘more expert’ than others. Different experts may have a preference for alternative cues that changes how each expert values individual cases, perhaps as a result of differing educations or experiences. For intra-expert variability, there are environmental factors such as time-of-day effects, changes in concentration, and fatigue that can alter how one expert perceives the same threat over time.

Another factor that has been shown to affect an expert’s reliability is that of previous incidents. Norman (1991) suggests that medical diagnostics is susceptible to bias introduced by recent cases. If a doctor has recently diagnosed a particular disorder, they often display a preference to diagnose this disorder again, and this has been shown to reduce accuracy by up to 20%. This is powerful evidence for the presence of the *Einstellung* effect, or mental set, first found by Luchins (1942). By consistently using a particular solution to solve a problem, when the nature of the problem changes so that an easier and more optimal solution exists, participants still apply the previous solution to the point that it inhibits their ability to now solve new problems. Mental set has been found to increase with stressors such as time-pressure and increased task demands (Luchins & Luchins, 1959).

Wiley (1998) first promoted the idea that domain knowledge could act as a mental set. She found that domain knowledge could inhibit creative problem solving tasks when a more general problem-solving approach was needed. She gave participants Remote Associate tasks (RATs - Mednick, 1962) in which three words are given (e.g., Plate, Blue, Cottage), and participants must find a word that is common to all three (e.g., Cheese). Wiley compared an ‘expert’ group that had a large amount of knowledge about baseball, with a ‘novice’ group that had a low amount of baseball knowledge. She found that, when a word-problem began with a term that had high associations with baseball (e.g., Plate), the expert group solved fewer RATs than the novice group that did not have this domain knowledge, showing that

domain knowledge can interfere with general problem solving. Bilalic, McLeod and Gobet (2008) studied the concept of expertise as mental set, and found that elite chess experts were still prone to demonstrating the Einstellung effect. Participants were presented with a series of chess-scenarios that had a familiar but sub-optimal move, and an optimal but unusual move. They found that experts often selected the sub-optimal move and reasoned that this was due to its familiarity gained over repeated previous exposure.

This research on mental set is critical for the foundations for this thesis that will make extensive use of the theory of mental set. Not only has mental set been established to inhibit general problem solving, but has also been shown to expand into the real world and inhibit problem solving there too. Given that stressors exacerbate mental, and these stressors are often found in the workplace, especially in high-risk military or security settings, then it may be more than possible that errors in these industries result from a form of mental set.

Wason (1960) found that participants tend to overcomplicate a relationship between cues when attempting to determine the relationship. He provided participants with three numbers (e.g., 2,4,6), and they were told the numbers followed a rule, and participants then had to determine what this rule was. In order to aid participants, they were allowed to generate number triplets themselves and given feedback on whether their triplets conformed to the rule or not. Although the actual rule was simple, it is reported that participants often neglected to identify it, instead favouring complicated and outlandish rules. This shows that participants who are making decisions when generating their own hypothesis tend to cling to the minutiae of details in the information presented, often to the detriment of the overall big picture.

1.2.2 *Fallibility in human decision-making*

A recurrent theme in decision-making research is that individuals often deviate from optimal and rational strategies or choices in systematic and measurable ways. It is widely believed that these deviations arise out of dual processes in reasoning. Dual process theories describe two systems of thinking; a heuristic ‘system 1’ process that occurs sub-consciously and without awareness, and an analytic ‘system 2’ process that is more complex in nature. Dual processes were first established by Wason and Evans (1975) after evaluating justification responses to the Selection Task (Wason, 1966). The selection task involved providing participants with a conditional sentence of the forms ‘if p then q ’ or ‘if p then *not*- q ’. Participants were then shown four cards; each card consisting of terms for p , *not*- p , q and *not*- q , and were told that each card had another term on the back of it. They were required to turn over the minimum number of cards needed to test the truth of the sentence.

Wason and Evans (1975) found that participants showed a ‘matching bias’ in terms of selecting the cards that were mentioned in the target sentence. For example, if the sentence was ‘If there is a K on one side of the card, there is a 7 on the other’, on being shown four cards H, K, 4 and 7, participants most often chose the K and the 7. The logically correct response is to select K and the 4, as these are the only two cards capable of falsifying the sentence. However, if the sentence was ‘if there is a K on one side of the card, there is *not* a 7 on the other’ then participants still showed a propensity to select the K and the 7, this time arriving at the correct response. Wason and Evans suggested that participants are using a heuristic of merely matching the terms in the sentence; this would create a correct response for the negation sentences, but fail to succeed on the affirmative sentences.

In addition to selecting cards, participants were also required to justify their reasons for which cards were selected. Wason and Evans found that participants attempted to rationalise their choices in terms of logic rather than acknowledging a matching bias. This

lead Wason and Evans to suggest the existence of a dual process, the selection of the cards was being done heuristically, the higher cognitive operation of explaining that choice being done analytically. More evidence of this rationalisation comes from demonstrations that participants will often rationalise and defend an incorrect solution even when shown contravening evidence (Wason & Johnson-Laird, 1972).

Tversky and Kahneman (1971) also describe occasions where systematic errors occur during judgement and decision-making. They showed that individuals do not correctly understand natural variation when dealing with small samples. They gave experienced psychologists questions about sample size and the conclusions that may be safely drawn from obtained data. Participants tended to ignore random sample variation in small cases and treat it as explainable, and a lot of descriptive narrative was attached to the initial data points.

To explain these findings, Kahneman and Tversky (1972) introduced the heuristic of *representativeness*. Similarly to the psychologists' errors regarding sample size, they found that many individuals commit errors when dealing with small samples, and that these errors tend to be based upon the extent to which the sample represents the population. For example, in a coin toss of six sequential cases, people estimate that HTHHTT (a sample with equal proportion of heads and tails) is more likely to occur than HHHHHT even though both patterns have the same probability. Kahneman and Tversky argued that this is because the first sample is more representative of natural frequency and thus associations are made which determine it more plausible to participants.

Another systematic error was reported in Tversky and Kahneman (1973a) based around *availability*. They found that people are more likely to assign higher probability and estimate a higher frequency of occurrence to events or activities that are more easily recalled or remembered. For example, individuals exposed to news items that contain violent imagery give higher estimates of crime and violence in the real world, than those that are not exposed

to news media of that type (Riddle, 2010). One explanation why availability, may influence and bias our interpretation that it is largely reliant on retrievability and imaginability (Tversky & Kahneman, 1974). Retrievability led to individuals to make overestimations about the size of a particular sample or population. For example, Tversky and Kahneman (1973a) gave two lists to participants that contained male and female names, and participants had to estimate whether there were more males or females on the lists. One of the lists contained the names of famous and identifiable males when compared to generic female names, and the other list contained famous and identifiable females as compared to generic male names. Participants tended to estimate more of the gender that was congruent with the identifiable names, even though both lists contained the same number of male and female names. Participants committed this error because highly retrievable and salient items on that list were more accessible, and so participants erroneously judged them to be more common.

The name lists created judgment errors because the names existed in participants' memory, but memory is not necessary for this fallacy to occur. The bias of imaginability allows errors of frequency without relevant domain knowledge. For example, the danger of an expedition may be judged on how many dangerous situations the participant can imagine. Instances where danger is available leads to overestimation of risk, but conversely, not considering potential likely dangers can lead to errors of underestimation. This may be true in our drone examples given so far. By operating in a military environment where the mission is to find and destroy the enemy, it may be that when given cues in the environment that suggest an enemy is present, it becomes easier to imagine a group of individuals being another group of enemies than it is to be a wedding party – which is a rare event by comparison and one in which operational military personnel are not actively looking for.

Stanovich and West (2000) coined dual processes System 1 and System 2. System 1 is the heuristic process, where information is considered unconsciously, implicit judgement

occurs, and is thought to be rapid. System 2 however is slow, deliberate, conscious, explicit and more analytic (Evans, 1989). Under this framework, availability, representativeness and matching are biases due to reliance upon system 1. Heuristics are often used in decision-making and problem solving to appear at an intuitive conclusion. Heuristics are shortcuts that individuals utilise, often based on experience. They can be efficient methods of solving quick and routine problems that leads support to them arising as an adaptive function (Gigerenzer & Gaissmaier, 2011). Gigerenzer and Gaissmaier found that use of heuristics often leads to ignoring some information; a judgement is arrived at without considering every available cue. They also found that in some circumstances this can still lead to accurate decisions where the burden of interpretation is low, but recognition is high.

When it comes to anomalies, heuristics would create difficulties in successful interpretation of an anomalous situation, because anomalies by definition are events that occur at a low frequency and whose appearance is unexpected and as such, may require additional cues be attended to in order to distinguish it from a separate event.

The availability heuristic predicts poor anomaly handling performance in practical settings. For example, a power plant has a malfunction detected through irregular sensor data (e.g., high temperature) anomalous and indicative of a threat. In this scenario, engineers must diagnose and rectify the problem. An increase in temperature is most frequently associated with using the wrong type of fuel. An engineer faced by this problem may attempt to diagnose the rise in temperature by generating hypotheses relating to fuel consumption as this is the most available explanation. An engineer familiar with this specific plant may be aware of a faulty section of pipeline that regularly bursts, leaking water necessary for cooling the system, and so a rise in temperature in one area may be due to a coolant leak in another area. The engineer that focuses upon fuel-centric solutions may never generate the correct hypothesis about the pipe without exposure to alternative relationships between the values

and variables in this example.

The impact of an availability heuristic is explored in the present study of anomaly handling, by examining how the frequency of a particular problem solution can bias solution choice when choosing from a sample of pre-selected hypotheses. For example, in our engineer example, while it would not be satisfactory to have novices innovate the notion of a broken coolant pipe when they have minimal knowledge of that domain, instead it would be better to provide potential participants with a number of solution hypotheses, including hypotheses based on fuel-type, but also including the broken pipe. This allows us to see under what conditions participants still commit decision errors when solution generation is not a barrier to success.

The representativeness heuristic is also important in the successful handling of anomalies, as it is important in determining that an event is indeed anomalous. If individuals are used to seeing or interpreting an environment in a particular way, then subtle yet critical changes in this environment may prove hard to detect as the anomalous environment may be more representative of an anomaly-free environment.

In order to examine the process of anomaly handling, a paradigm must be generated in which successful interpretation differs from a failed one, and must also maintain a mechanism that allows us to detect the use of inappropriate heuristics. Our system (described in more detail in Chapter 2) comprises a variable that represents the supply of a resource distributed around a number of nodes, and the distribution of that variable is determined by other factors. The system is one of intelligence in a conflict setting; supply of weapons act as the resource variable, different population zones act as the nodes, and local environmental conditions determine the available supply and distribution of the weapons. In this case, it is 'normal' to have weapons distributed equally to all zones, and an accumulation of weapons in one zone suggests an anomalous event. One common interpretation of this accumulation of

weapons is that a conflict is impending. It is the job of the intelligence analyst to detect the changes in supply that suggest conflict. The representativeness heuristic can be examined by the introduction of anomalies, these would be fundamentally different events occurring within the simulation, but that appear similar to a conflict. In principal, this will mimic the difference in our drone example, between a meeting of insurgents and a wedding party, two events that appear similar enough to fit the description of a ‘signature strike’ yet result from two extremely different causes. In an example from our system, adverse weather conditions may conspire to prevent delivery of weapons to particular locations, where the surplus is then redirected to remaining accessible locations, this combination of events then results in a weapons distribution that looks like an impending attack. In this case the correct solution would not be one of identifying an attack, but by using additional (and perhaps often ignorable) information, perhaps through system 2 reasoning, to determine the true cause of the event.

The system used in the current research is essentially a causal reasoning task. The user must determine the cause of the resulting distribution as either hostile (weapons stockpiling for impending attack) or no threat (weapons surplus due to supply-side issues) based on the input of the system, as the outcomes of both may appear identical.

1.3 From laboratory to field – Human involvement in anomaly handling

1.3.1 Anomalies in naturalistic environments

Despite the importance of determining fundamental factors that underpin success in anomaly handling tasks in the laboratory, it still remains our aim to explain how this process can be important in understanding performance in naturalistic environments. Anomalies are often indicative of an undetected fault or threat, and by successfully interpreting these accurately and quickly it may be possible to mitigate potential damage to the systems in

which these anomalies are experienced. Research suggests that human error plays a large role in the failure to detect anomalies in industry. It has been estimated that between 70-80% of incidents in high-risk industry occur due to human error (Reason, 1990). With the consequences of such incidents being potentially catastrophic, the role of this human error has provided an attractive topic of psychological investigation. Sneddon, Mearns and Flin (2006) conducted research into the cause of such potentially catastrophic incidents on offshore oilrigs and found that attention was a large cause of human error. In this industry, incidents can accrue huge loss of life, large monetary damage to assets, and grave harm to the environment (Cullen, 1990). Another significant conclusion of Sneddon et al.'s work was that a critical component of human error was a poor understanding or conceptualisation of the workspace.

The dangers of a poor understanding of the workspace can manifest into potential incidents in a number of ways. For example, poor conceptualisation of a problem could lead to inefficient decision-making that fails to provide an adequate solution to deal with a target problem. Alternatively, if the environment in which a problem is encountered is poorly understood then it could be that anomalies in the workspace are not attended to, or noticed as being indicative of a threat, and do not become a recognised danger. As Klayman (1988) noted with respect to cue identification, if a work environment is poorly understood, then an individual's ability to determine useful cues from that environment is compromised.

A good example of the latter is the case of the nuclear incident at Three Mile Island in the United States in 1979. The incident developed when a valve malfunctioned, and failed to close as it should have done. Although the problem was mechanical in nature, there were a series of human errors that exacerbated the situation. According to an official Nuclear Regulatory Commission report (Rogovin, 1980) the fault with the valve was not discovered by the operators for hours after the malfunction. This was due to two main reasons; the first

was that operators believed a signal on the control panel told them that the valve was closed. In fact the signal actually represented a different but proximate mechanical component, and over time, operators had habitually incorporated the function of this component as a proxy measure for the state of the valve. Second, additional information available to the operators such as temperatures and pressures at various parts in the system would have been diagnostic of the open valve; however, because no problem was thought to exist, this was not attended to. It could be argued that both of these failures in human reasoning arose out of a poorly understood conceptualisation of the workspace. If the operators had better understood the roles of the signals, then they may have not ruled out a problem with the valve. If they had attended to the anomalous additional information, then they may also have realised the presence of a problem.

So why couldn't the operators at Three Mile Island diagnose the problem? The simple answer was that they didn't realise a problem was occurring. Smith (1989) differentiates *problem identification* from *problem detection*. Problem identification is the series of behaviours or cognitive processes that are executed when a problem exists, and the cause is aimed to be determined. Problem detection is the arousal of a suspicion that something may be wrong with a system that deviates from its normal state of operation, which then allows problem identification to occur. Whilst the Three Mile Island operators would have no doubt been able to diagnose the problem if tasked to, their belief that the plant was operationally functional meant they had no need to. This is a case of a failure of problem detection.

Problem detection was first described by Cowen (1986), who provided a rudimentary explanation as to how it occurred. Cowen believed that problems were detected when variations between what was happening and what we want to happen accumulate to pass a particular threshold that allows us to notice something is not optimal. Whilst this discrepancy-accumulation model was the first to tackle problem detection, and provides one

route where problem detection may occur, it fails to provide a fully substantiated model that accounts for all instances of problem detection.

The issue of problem detection was further explored by Klein, Pliske, Crandall and Woods (2005). They attempted to create a framework of problem detection in a more flexible manner than Cowen's discrepancy accumulation model. They elaborate on a number of cases of critical incident decision-making where the problem detection itself was a prominent feature. By using exemplar cases, they show that Cowen's model is not enough to account for problem detection. One exemplar case in question comes from an experienced paediatric nurse who immediately noticed something wrong with an infant and followed it up, ending in the successful diagnosis of a problem when the same symptoms were available to a lesser experienced nurse who failed to detect the problem (Crandall & Getchell-Reiter, 1993). The case shows that problem detection need not be an accumulation of discrepancies, as the nurse noticed a problem immediately. Also, it shows that an element of expertise and experience must be accounted for in problem detection, as the exact same information had been non-indicative of a problem to a more junior member of nursing staff.

Klein et al. (2005) argue for two additions to the discrepancy-accumulation model. First, a discrepancy is a violation of expectation rather than a deviance from a goal state. This addition provides an explanation as to how individuals of differing experience can have differences in ability to detect problems, as their expectations alter with expertise and allow a finer resolution with which to detect problems. For example, in the nurse exemplar above, while the junior nurse may know that the colour of baby is indicative of its health (blue is unhealthy, pink is healthy), the senior nurse has a higher sensitivity and can better discriminate at what point the baby's colour indicates a problem. The second change is that accumulation is not necessary, as single salient cues may be enough to trigger problem detection, and also a number of discrepancies can be explained away that do not result in

problem detection.

Problem detection is important for anomaly handling, because some anomalies may appear similar to a normal situation, and in order to interpret these anomalies successfully, the observer must first differentiate that a change has occurred. The nature of problem detection was incorporated into our simulation. This was possible by making some problems solvable using a single cue with different characteristics. Our anomalies were then created to have similar characteristics to these distinctive cues, but where this information must be combined with another cue to successfully solve. If our decision-makers are using heuristics for the majority of problems (which are not anomalies) then problem detection will become an issue for anomaly handling. This should be reflected in reduced participant performance on our task.

Klein (1989; Klein, Orasanu, Calderwood & Zsombok, 1993) described problem detection in his Recognition Primed Decision Making model (RPDM). RPDM is characterised by comparing a new problem to an internal repertoire of previously encountered problems in order to classify and arrange a suitable response to an emergent problem. The RPDM model works on a series of problem reconceptualisations that arise when an expectation is violated. Expertise in the nurse exemplar, above, may come from either recognition of the problem from having experienced similar cases previously, or from a violation of the expectation that the child is healthy as the child did not match examples of previously identified healthy babies. The more junior nurse was not able to detect the problem perhaps because she had never encountered this particular set of symptoms as so could not recognise it as a trajectory to serious illness.

When it comes to successful interpretation of anomalies or discrepancies, the RPDM may not be capable of such a task, which could result in the failure to solve a problem when the problem looks like previously encountered problems. For example, RPDM works off a

library of previous experiences that offers a repertoire of signs, symptoms, classifications and solutions. If a problem has not been seen before, then it may still be anomalous enough to arouse suspicion and violate expectation, but it will not trigger any of the cues to stored solution knowledge. If we classify the functionality of a system and possible problems based on a library of prior experience, then we do not allow room for innovation, and may even encourage fixation on non-optimal decision paths (De Keyser & Woods, 1993). For example, if operators at Three Mile Island had noticed that a signal was present that previously had meant the plant was functioning as expected then there would be no reason to attempt to innovate a solution to the problem of detecting the open valve.

A review of problem detection (Klein et. al. 2005) suggested a role for sense-making in the ability to interpret anomalies. They explain that anomalies are actually frequent events in many industries, and that deviation alone is not enough to classify a problem. Instead, we have a conceptualisation of the workspace that can be used to explain away anomalies. If a particular conceptualisation doesn't explain the anomalies, then a new conceptualisation is recreated, until all anomalies are accounted for and there is no longer a problem.

Sense-making is elaborated in the data-frame model (Klein, Moon, & Hoffman, 2006). In this model, the individual possesses a starting state based upon the data available to them. A 'frame' is constructed which fits these data and acts as the individual's conceptualisation of the workspace. As data change in real-time, a frame may be compromised when anomalies are introduced – the individual may try and preserve the frame by tracking the anomalies, detecting further inconsistencies, or judging the plausibility or quality of the data. The frame can then be elaborated and questioned, new data can be sought out and inferences can be drawn upon newly discovered relationships in the data. This ends with a reframing of the workspace, that encompasses all of the previous anomalies and thus removing them from the new frame (Klein, Philips, Rall, & Peluso, 2007). To put this into

context, we can return to our opening example of drone strikes. When an operator is assessing surveillance materials, they possess a 'frame' of their workspace in which they are looking for hostiles, when the operator comes across a large group of males this may be consistent with that frame. By further monitoring the group, the operator may track those individuals trekking cross-country without weapons. The lack of weapons may be unusual in this case, and be incompatible with the 'frame' that these people are insurgents. As such, the operators frame adjusts to incorporate this new information. Perhaps the group is travelling unarmed to avoid suspicion of ground forces, perhaps the weapons are concealed, or perhaps the group are not hostiles at all. The operators 'frame' switches to one in which the group may or may not be hostile. Continued monitoring of the group shows that they turn up at a religious building where traditional marriage clothing can now be seen on the individuals, and a ceremony takes place. The operators 'frame' now shifts again to one in which the operator understands that what they are viewing is a wedding and not a group of hostiles, and so no strike is conducted.

The addition of sense-making to anomaly handling enriches accounts of the causes of human error (Sneddon, Mearns, & Flin, 2006). A poorly conceptualised workspace may be the cause of human error because of a lack of successful sense-making. Support for this idea comes from research by Malakis and Kontogiannis (2012), who found that highly salient concepts of successful problem detection in air traffic controllers had significant overlap with traditional concepts in sense-making such as situational awareness.

Many studies looking at problem detection in particular domains have been conducted examining interview responses gained after using the critical decision method of cognitive task analysis (Klein, Calderwood, & MacGregor, 1989; Hoffman, Crandall, & Shadbolt, 1998) and then trawled for examples when problem detection was a main feature of the critical incident. Research has also been done on problem detection in a naturalistic setting.

For example, Mumaw, Roth, Vicente and Burns (2000) conducted naturalistic observations inside a nuclear power plant. Rather than examining historical critical incidents and looking for occasions when detection was flagged as a prominent feature (which must be vulnerable to issues involving memory), Mumaw et al. conducted observations while the power plant was in a functionally operational condition. This allowed real-time monitoring of problem detection as it occurred. It also allowed conclusions about the vigilance, strategies and procedures that staff used before the problems were detected.

Findings from both Mumaw et al. (2000) and Klein et al. (2005) converge on similar issues surrounding problem detection. Research from multiple domains has shown that the ability to interpret anomalies is not a task that occurs independently from the ability to detect anomalies. Nor is the conceptualisation or framing of a workspace a task that is accomplished easily. Both findings show that problem detection usually arises out of the presence of anomalies. But industry often requires the individuals to function in an environment where anomalies are commonplace. Mumaw et al. (2000) observed that for power plant operators, signals often came in the form of alarms, and that there were a vast amount of signals that each operator was responsible for. When monitoring the operational functionality of the power plant, many alarms would sound, some continuously. So the issue isn't how to detect these alarms from a stable, functional state, but how to interpret patterns of anomalies that could lead to potentially critical incidents from a noisy and anomaly filled background. The process of problem detection goes more than just a violation of expectation, but becomes a process of anomaly handling.

We use the term anomaly handling for the approach by which information is processed by decision makers in the detection, identification and resolution of potential problems, specifically when those problems present with a similar appearance to routine problems that are more frequent and require a different solution. So while problem detection

is a feature of anomaly handling, the latter also encompasses the ability to infer the health of a system from those initial anomalies.

While this research summarises that people are capable of identifying and tending to anomalies, it does nothing to determine the condition, factors and boundaries at which success can be achieved. Post-hoc descriptions of a decision are excellent ways to create frameworks that describe this process, but to understand the limits of human anomaly handling it would be more appropriate to create a paradigm that can be utilised under laboratory conditions, yet simulates a real world experience. This way, it is possible to manipulate factors and better understand the conditions that lead to anomaly-handling success.

2 Developing a simulation for exploring anomaly handling

2.1 An introduction to the experimental paradigm

2.1.1 Overview and goal of the simulation

In order to develop a simulation to explore decision-making in a controlled but realistic way, a simulated environment was designed to be simple enough to allow individuals to understand the context of the environment and to be able to form expectations and create predictions of how the system would behave. The psychological literature on cues, their acquisition from an information environment, and the limits to which experience would not inhibit performance were all taken into consideration during the development of the task. As a direct result of this research, the design was informed by this previous research in a number of ways. First, the number of available cues was limited so as to give novices an ability to comprehend and participate without memory or domain knowledge acting as a barrier to performance. Second, the paradigm was designed so that solutions to some types of problem would heavily favour a single cue more than other cues. This was to create an optimum environment in which participants may develop heuristic strategies by placing a heavier weighting or importance on a single cue. As an extension to this, a set of problems (representing out anomalies) were designed in such a way that reliance on a single cue for solution generation would create a tell-tale series of errors that would identify the use of such heuristic strategies.

The chosen paradigm had to be based upon a relatively simple rule-set to be accessible to naïve participants, yet be sophisticated enough to allow complex interactions within the system to arise, these complex interactions will represent our anomalies.

The context chosen for our environment was that of intelligence analysis within a conflict setting, where each participant would play the role of an intelligence analyst. The simulation represents the security picture within a fictional country. The objective of the

participant is to make a security assessment about the state of the country. During the simulation, weapons flow between different locations and participants will have to determine the underlying reasons for the movement of weapons. The simulation offers two possible explanations for changes in arms traffic; this is either due to an impending conflict within the country, or because the supply of arms is being affected by the environmental conditions. Participants will have to use multiple information channels within the simulation to decide which explanation better fits the data available to them.

2.1.2 *Information available within the simulation*

Within the simulation, participants have access to a map of the country. Marked on the map are four different zones. Also on the map are the known supply routes by which arms are trafficked into each zone. Finally, the map also contains the types of environmental hazards that each supply route must cross. This map and all relevant information can be seen in Figure 1. While a number of trials will be given to each participant, the map remains consistent throughout the experiment. A trial consists of two sets of data representing weapon volumes and environmental conditions.

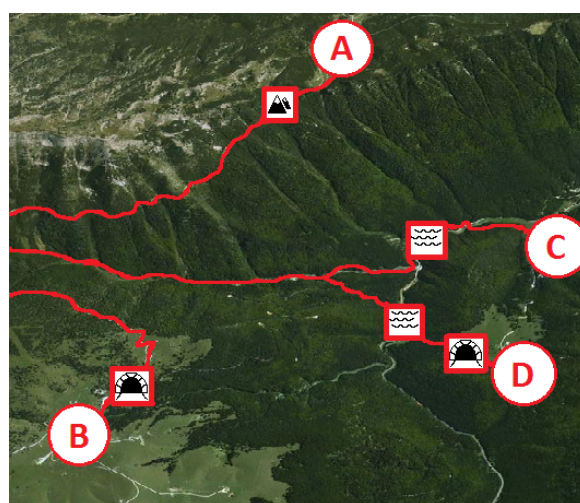





Figure 2.1. *Map of zones, supply routes, and environmental hazards used in the simulation.*

Within the map, four zones can be found, labelled A-D. Red lines represent the supply routes by which arms are trafficked into each zone. The map also contains the type of environmental hazards, with a unique icon for each type of hazard. The types of hazard are mountains , underground tunnel , and a river .

Each participant also has access to two tables of information. These tables represent the number of weapons in each zone and the local environmental conditions, each over five time-points. Examples of these can be found in Figure 2 below. Each trial consisted of one table of each type

Conflict Indicator - Estimated Weapons				Local Environmental Conditions		
Zone A	Zone B	Zone C	Zone D	Rainfall (mm)	Temperature (°C)	Seismic Activity
398	399	407	394	3	12	2.1
406	395	396	394	9	18	2.1
393	408	397	397	2	19	2.7
399	406	406	397	8	17	2.3
398	395	392	408	10	19	2.9

Figure 2.2. *Examples of weapon information (left), and environmental information (right)*

Within the simulation, each environmental condition is related to an environmental hazard. A reminder of this information is also contained within the simulation to not make memory a reason for failure, see Figure 3 below. The interactions between environmental conditions and hazards will be explained further in the technical section.

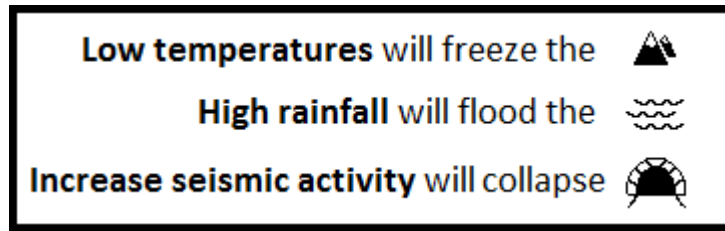


Figure 2.3. *Reminder of environmental / hazard interactions available throughout experiment.*

Finally, the last object within the simulation was an on-screen prompt that captured the decision of the participant. This was an on-screen box that relayed the options to the participant, alongside a description of which key to press for each option (Figure 4.)

Action status: Report needed					
Please press the key that best describes what is happening					
Threat detected - conflict occurring in:				No threats detected	
Zone A	Zone B	Zone C	Zone D	No change in arms traffic detected	Increased arms traffic due to environment
A	B	C	D	P	T

Figure 2.4. *Decision box, explains which key to use to register each possible outcome.*

When combining each of these elements together, Figure 5 shows a typical representation of a single trial, and what a participant sees within the simulation. Between trials, only the two tables containing data would change.

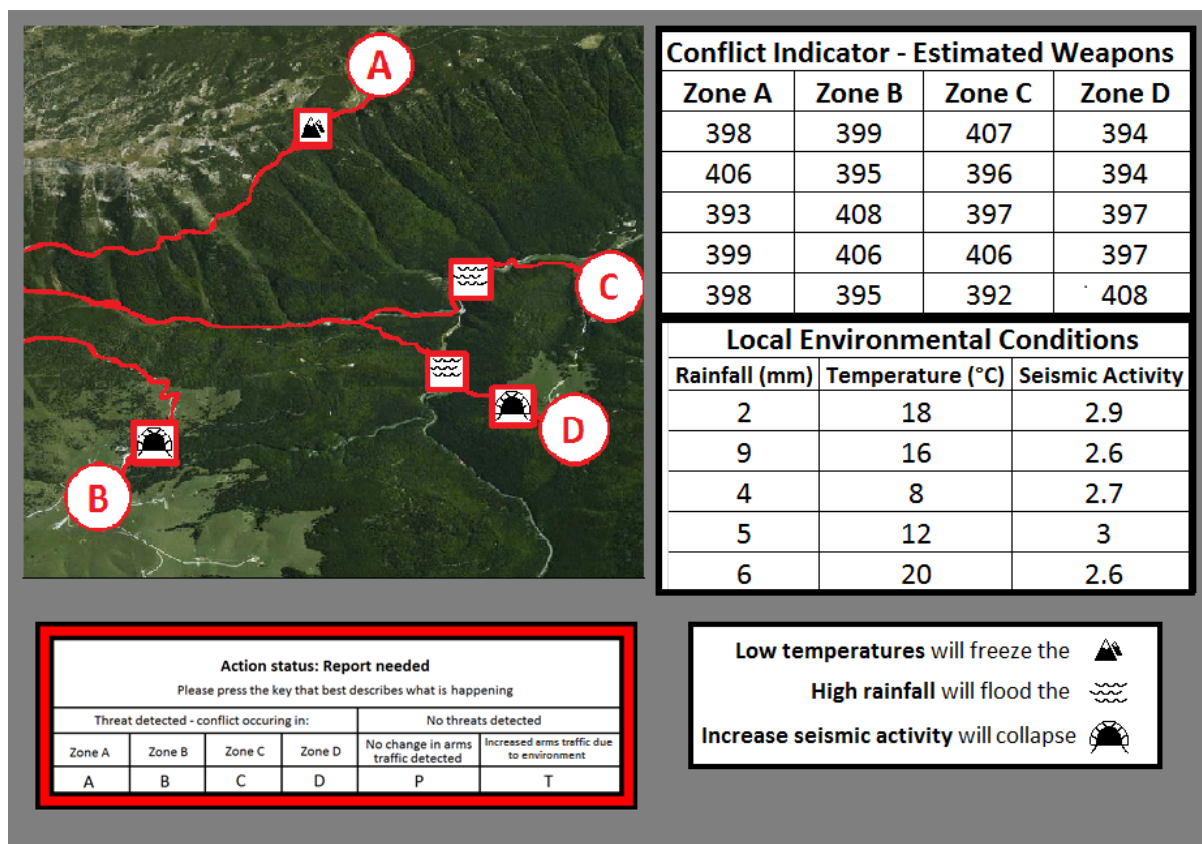


Figure 2.5. Typical representation of a single trial within the simulation.

2.2 Technical rules overview

2.2.1 Environmental events

Within the simulation a number of environmental events are possible. Each environmental event is represented by a rapid change in a corresponding environmental variable. An increase in rainfall will flood the river, a decrease in temperature causes the mountain pass to freeze, and an increase in Richter scale recordings is indicative of recent earthquakes that collapse underground tunnels. A reminder of these interactions is present within the simulation at all times for participants to access. It is possible to determine which zones are affected by each event, by examining which supply routes cross affected hazards. If

a supply route crosses a hazard that has been negatively affected, that supply route is deemed treacherous. Treacherous supply routes are only capable of delivering half (50%) of the normal supply of weapons. All undelivered arms due to treacherous supply routes are distributed across all remaining non-treacherous, safe supply routes. For an illustrated example of this, please see Figure 6 below.



Figure 2.6. *Example of normal (left) and treacherous (right) temperature data and distribution*

To deliver arms to Zone A, the mountain must be crossed (as indicated by the red supply line going to Zone A crossing the icon that represents the mountain). It is explained that low temperatures will affect the mountain. In the above example, on the left the temperature remains normal over the five time-points, and the weapons can be supplied as the hazard is not treacherous. This leads to a ‘normal’ amount of weapons being delivered to Zone A, and none are redistributed (as indicated in Figure 6 by 100% supply to Zone A and 0% return rate. In the example on the right, we have an extreme drop in temperature over time, from 15 to -22 degrees Celsius; this makes the mountain treacherous. As a result, only half (50%) of the weapons can get through to Zone A. The remainder (50%) are sent back for redistribution to other non-treacherous zones (not pictured here) assuming all other zones are safe, these weapons would then be supplied to Zones B, C and D alongside their normal supply, increasing weapons supplied in these other zones

Changes in distribution of weapons can be monitored in the numeric weapons table.

At the start of the simulation, the distribution of weapons across all nodes is equal. Each node receives the same number of weapons and this distribution is consistent over time. This number only changes when supply routes are being affected by environmental events, or in the case of a conflict. If there are no environmental events, the distribution of weapons will remain equal across all time-points. Figure 7 is an example of a ‘no change’ weapons distribution where no events are present. It is important to note that even in the ‘no change’ trials, there is some small variation between time periods (time data is vertical, top is the earliest data, with the bottom row representing the most recent). However, these small changes are not large enough to constitute an event, either environmental or conflict.

Conflict Indicator - Estimated Weapons			
Zone A	Zone B	Zone C	Zone D
404	399	405	393
393	408	394	392
401	396	399	403
404	402	407	403
404	403	397	401

Figure 2.7. *Example of a ‘no change’ weapons distribution.*

If there are environmental events, the distribution of weapons changes to reflect the rules above on treacherous hazards. This results in the number of weapons decreasing in zones using treacherous supply routes, and increasing in all remaining zones. In this case, the presence of environmental events drives the change in weapons distribution in a causal manner. An example of this may be seen in an extension of our earlier example of the treacherous mountain. If the mountain is treacherous (which must be crossed to supply Zone A), then Zone A will only receive 50% of its intended weapons, with the remaining amount being redistributed. An example weapon table for this can be found below in Figure 8.

Zone A	Zone B	Zone C	Zone D
400	394	396	400
397	392	405	397
329	455	453	441
255	464	451	463
200	470	471	476

Figure 2.8. *If the mountain was to become treacherous, the above weapons distribution would be seen. Note that weapons in Zone A have reduced over time to half of the original number. The remainder of this is then split between all remaining safe Zones. This assumes that all other zones are not treacherous.*

A further example of how environmental conditions affects weapons supply can be seen in Figure 9. In this example, we can assume the rainfall has critically increased. We know that this will make the river treacherous. As per the map, the river must be crossed to supply Zone C and Zone D. The result is that weapon volume in Zone C and Zone D reduces to half its original number, and that the difference can be found in the increases to Zone A and Zone B.

Conflict Indicator - Estimated Weapons			
Zone A	Zone B	Zone C	Zone D
398	400	392	404
395	400	405	401
464	472	335	332
555	557	255	255
588	594	199	202

Figure 2.9. *Example of a weapon distribution table in the result of a flooded river.*

When the changes in weapons can be fully explained by the presence of environmental events then the system state is one where there is no threat, and the changes in arms traffic is due to the environment. This is not indicative of a conflict.

2.2.2 Conflict events

Conflict events are a separate type of event that is distinct from environmental events. Similar to the environmental event, the presence of a conflict affects the supply of weapons to

zones. However, unlike environmental events, there is no indicator that states the presence or absence of a conflict, and this event must be inferred from the distribution of weapons combined with the absence of appropriate environment conditions to account for the resultant distribution. If there is a conflict, rules state that the weapons distributed to all non-conflict zones is decreased by half (50%), creating a surplus to be redistributed to the conflict zone. See figure 10 below for an example of a conflict seen via the weapon distribution table and the environmental conditions.

Conflict Indicator - Estimated Weapons				Local Environmental Conditions		
Zone A	Zone B	Zone C	Zone D	Rainfall (mm)	Temperature (°C)	Seismic Activity
392	406	398	396	6	8	2.4
396	396	405	405	2	18	2.9
324	605	324	328	10	18	2.3
251	816	253	251	4	12	2.4
202	1002	200	197	4	20	2.5

Figure 2.10. *Data tables for a conflict. The weapons data (left) shows a large increase in Zone B, and all other zones have reduced by 50%. The environmental data (right) shows no significant changes that would trigger a hazard. With the lack of environmental explanation, the change in weapons must be due to a conflict in Zone B.*

A conflict event creates a large increase in weapons in the zone of origin, and lowers the number of weapons available across all other zones. Environmental events differ by creating a decrease in the number of weapons available to zones that are deemed treacherous (assigned by the supply route crossing an affected hazard), and providing all non-treacherous zones with an increase that is dependent upon the number of treacherous zones. It should be noted that the increase in weapons caused by a conflict event is always greater than those caused by a single environmental event.

2.2.3 *Adding complexity*

The previous examples outline the simple distribution changes that conflict and environmental events incur. These are the simpler problems. High solution rates on these problems show that participants understand the rules enough to engage in the simulation. To reflect the nature of these problems, they have been named; no event, simple environmental event, and simple conflict event respectively. In each of these, there is either no event, or one event occurring at any given time, making the distribution follow the rules clearly.

There are two types of problem that are special instances and represent a more challenging problem that requires additional inferences to be made before solving correctly. These two problem types are complex conflict events, and complex environmental events. Complex conflict events problems possess a conflict event, and an environmental event that affects non-conflict zones. In this problem, there are cues for both the conflict system state, and the environmental system state. However, the environmental cues are not relevant and the distribution can only be explained by the conflict event, as an environmental event alone cannot cause the level of change observed. An example of this can be found in Figure 11 below where the weapons table (left) shows a large increase in Zone B, and decreases in all other Zones. However, there is still activity in the environment table (right) where we can see a large increase in rainfall. This would flood the river, and makes Zone C and D treacherous. However, if rainfall were the only event, both Zones A and B would increase, whereas Zone A has seen a reduction. The only explanation based on the supply rules given is that there is a conflict in Zone B at the same time as an environmental event. This will also explain why Zones C and C are lower than Zone A. Zones A, B and C have reduced as a result of the conflict in B, and Zones C and D have reduced even further due to the treacherous hazard. In the role of intelligence analyst, participants are told their primary function is to identify and spot conflicts. In cases where there is both an environmental event, and a conflict,

participants are told that identifying the conflict takes priority – and so stating there is a threat due to conflict is the correct answer.

Conflict Indicator - Estimated Weapons				Local Environmental Conditions		
Zone A	Zone B	Zone C	Zone D	Rainfall (mm)	Temperature (°C)	Seismic Activity
399	403	392	403	8	13	2.9
392	398	402	407	4	17	2.9
327	602	246	248	15	17	2.9
247	1004	203	202	37	12	2.1
201	1122	99	99	64	20	2.1

Figure 2.11. *Example weapons and environmental tables from a complex conflict event*

The second complex problem is the complex environmental event. In these problems there are two environmental events. These two events block three of the four possible supply routes. This causes a 50% decrease in each blocked zone, with a resulting 150% increase in the only remaining zone with safe supply routes. While this change can be explained by the environmental events, the final weapon distribution is identical to those of the simple conflict event problems. In this case, evidence for both conflict and environmental system states are present, but the changes observed are not large enough to meet the criteria if a conflict were also present, leaving the environmental event as the only possible correct solution. Figure 12 shows example data tables of this scenario below. In this example, the weapons table (left) shows a large single increase in Zone B, and all other zones have reduced by half. This matches the type of data seen in a simple conflict event, however, when we examine the environmental table (right), we can see there are two environmental events occurring simultaneously. The low temperature would create a hazard in the mountain, effecting Zone A, and the high rainfall would create a hazard in the river, effecting Zones C and D. In this case, Zones A, C and D are all treacherous. Following the supply rules set out, treacherous routes have their supply halved, with the remainder evenly divided among all remaining non-

treacherous zones. In this case, Zone B is the only non-treacherous zone and so receives all of the surplus weapons itself. This example can be fully explained by the presence of environmental events only, and so no threat needs to be recorded.

Conflict Indicator - Estimated Weapons				Local Environmental Conditions		
Zone A	Zone B	Zone C	Zone D	Rainfall (mm)	Temperature (°C)	Seismic Activity
400	395	407	406	3	15	2.2
405	402	404	407	9	10	2.7
327	609	329	331	20	-3	2.8
251	784	253	251	36	-14	2.9
204	1019	203	201	62	-25	2.2

Figure 2.12. *Example data tables of a complex environmental event*

Instructions given on the explanation and rules of how the task is to be completed can be found in Appendix 1.

2.2.4 Inferences

During this task, there are a number of inferences that must be made for the participant to achieve the correct solution. These inferences can be divided into two phases. In the first phase, the participant must identify any potential treacherous hazards caused by the environmental conditions. Participants are told that rapid change in environmental variables cause environmental events which make hazards treacherous. Participants must make judgements as to which point conditions become critical. Rainfall is normally 0-10mm, but rises to 70 when critical. Temperature is normally 10-20C but drops to -20C when critical. Richter values are normally 2-3, and rise to 6 when critical. While the temperature and rainfall metrics represent a distinct change over a larger range of values, the Richter scale is much more sensitive with a shorter range of values, and participants will need to learn to understand the difference between these ranges to successfully understand the causes of each

different environmental event. Environmental events can be inferred by observing the critical threshold values being exceeded. To help establish these thresholds, environmental events create extremely distinctive weapon distribution profiles that can only be associated with a particular metric. For example, at the introduction of the first environmental event caused by an increase in Richter value, the weapon distribution changes so that collapsed underground tunnels are the only possible explanation. At this point, participants can establish a baseline for what an exceeded threshold looks like for that variable.

Once participants have inferred the state of the hazards from the environmental conditions, a second inference must be made about how that is expected to change the distribution of weapons. If we assume that there is no error in the ability to determine the presence of treacherous hazards, then distributions should behave accordingly when experiencing each of these events. Table 2 summarises the weapon distribution changes for all possible combinations of environmental and conflict events.

Table 2. *List of possible scenario events for both complexity (simple and complex) and event type (environmental and conflict) showing their resulting change in weapons distribution in each zone.*

	Event	Treacherous	ΔA	ΔB	ΔC	ΔD
Simple	Low temp	A	-50%	+17%	+17	+17%
	High rain	CD	+50%	+50%	-50%	-50%
	Seismic	BD	+50%	-50%	+50%	-50%
	Conflict A	--	+150%	-50%	-50%	-50%
	Conflict B	--	-50%	+150%	-50%	-50%
	Conflict C	--	-50%	-50%	+150%	-50%
	Conflict D	--	-50%	-50%	-50%	+150%
Complex	Low temp + high rain	ACD	-50%	+150%	-50%	-50%
	Low temp + seismic	ABD	-50%	-50%	+150%	-50%
	High rain + seismic	BCD	+150%	-50%	-50%	-50%
	Conflict A + high rain	BCD	+200%	-50%	-75%	-75%
	Conflict A + seismic	BCD	+200%	-75%	-50%	-75%
	Conflict B + high rain	ACD	-50%	+200%	-75%	-75%
	Conflict B + low temp	ACD	-75%	+175%	-50%	-50%
	Conflict C + low temp	ABD	-75%	-50%	+175%	-50%
	Conflict C + seismic	ABD	-50%	-75%	+200%	-75%
	Conflict D + low temp	ABC	-75%	-50%	-50%	+175%

Each event and combination of events has unique profile except those of the simple conflict events and the complex environmental event. These two problem types have identical distribution profiles but can still be solved correctly. The complex environmental events will have the appropriate environmental metrics exceed their critical threshold, whereas the simple conflict events have normal environmental data. It may be expected that the two may be incorrectly identified if participants use heuristics to complete the assessment based on the weapons data alone. However, this misattribution bears no basis in logic. For example, the

complex environmental event of Low temp + high rain creates treacherous hazards between zones A,C, and D, creating a weapons distribution similar to a conflict in Zone B. It is expected that participants may fail to identify this difference and incorrectly select a conflict event as the solution if they are not using logical reasoning, and are instead using heuristics. The table below demonstrates why that cannot be the case. If a conflict in B were to occur at the same time as the environmental events, then the change we see should be much more pronounced. It is not, and so there cannot be a conflict occurring at this time. This leaves the only remaining cause of the observed change to be caused by the environmental conditions.

Table 3. *Table showing similarity in weapon distribution of simple conflict and complex environmental problem types with additional outcome for both combined.*

	Event	Blocked	ΔA	ΔB	ΔC	ΔD
Simple Conflict	Conflict B	ACD	-50%	+150%	-50%	-50%
Complex Environmental	Low temp + high rain	ACD	-50%	+150%	-50%	-50%
Complex Environmental + Conflict	Low temp + high rain + conflict B	ACD	-75%	+225%	-75%	-75%

Similarly, it is possible to theorise that participants may mistakenly categorise the complex conflict events as environmental events. While it is true that an environmental event is occurring, it is not possible for this event alone to be the cause of the observed change, and can only be explained by the presence of a conflict. Because of these interactions it is possible for each scenario to have a conclusive correct answer that can be determined from

the data being provided, while at the same time leaving room for systematic errors to be made if participants use heuristics in the form of reducing the number of cues they use, or rely on recognition-based pattern-matching.

2.3 The use of simulations in psychological research

Decision-making researchers have often used simulations to explore specific domains. For example, Hilburn, Jorna, Byrne, and Parasuraman (1997) studied the decision-making skills used by air traffic controllers in a simulation of an air-traffic control (ATC) workspace. The information available mirrored the type of information output to actual ATC displays. The method of problem-solving was also as similar to the actual job, including the response options available to the participants, who were experienced air traffic controllers. The aim of the research was to gain quantitative information about how ATC controllers organise, construct, and carry out their decisions. Similarly, Omodei and Wearing (1995) created *Fire Chief*, a simulation of a firefighting command response system to assess the structure of decisions made by fire commanders as to where to deploy available resources. Although this simulation was more abstract than Hilburn, et al. (1997), the information used to inform each decision, such as the location of a fire, the size of a fire, and current weather trends were all available, mirroring the critical information available in the real world, alongside realistic response options.

A review of simulations used to conduct psychological research was carried out by Gonzalez, Vanyukov and Martin (2005), who found that simulations, such as Omodei and Wearing's *FireChief*, are rated high enough in terms of complexity and fidelity that they are to be considered a powerful tool in the study of complex decision making. They argue that such simulations allow researchers to collect quantitative data regarding performance that may remain undetected with qualitative research methods such as naturalistic decision making, and that they are a middle ground between laboratory studies where task

performance is often not indicative of expertise, and field studies, whose research methods do not allow a degree of experimental control.

In both Omodei and Wearing's *Fire Chief* and Hilburn et al.'s ATC scenario, the simulations are used to create environments for studying decision-making that attempt to be an accurate reflection of real systems in order to better account for real-world experience. In such systems, an experienced participant is expected to perform better than a novice because they can rely on their domain knowledge and strategic experience to formulate a response that will lead to a predictably better outcome. For example, Johnson et al. (2014) examined the decisions to use deadly force in a shoot/no-shoot task within a simulated video environment, comparing military personnel and police officers, and novices, who had no formal deadly force training. They found that experts were more likely to make a correct decision to use deadly force than novices. The simulation was of sufficient fidelity to facilitate transfer of real world experience, and captured performance differences between individuals of different ability.

The critical factors of a simulation that allow it to be used successfully for psychological research are fidelity and complexity. A simulation must have a high fidelity to the situation in which it is trying to mimic. This includes having the same critical information available, and replicating the same environments that are possible with the real task. This is to allow the transfer of skills and knowledge from the real environment to the simulation. If the simulation is not high enough in fidelity, then an expert in the real world would need to 're-learn' the simulation as their skills and experience would not carry over. For example, one prolific use of simulations for psychological research is that of driving performance, such as the effect of emotional phone conversations on driving (Briggs, Hole, & Land, 2011). In driving simulators, not every detail needs to be the same as in a real car. For example, many driving simulators used in psychological research will not have foot pedals or a steering

wheel. In these experiments, the most important part of the task is what a driver can see – and so the simulators cater to creating a visual experience similar to real driving. As such, our simulator includes only those aspects of military intelligence required for the task assessment to be conducted with certainty. These are the combined elements of our simulation including the geography, and the data streams.

The second element for a good simulator is that of complexity. The more complex a simulator is then the more situations it is capable of handling, and the more nuanced results any input can have. The two elements of complexity and fidelity go together, with an increase in complexity usually facilitating a higher fidelity. For example, in Omodei and Wearing's *Fire Chief*, although it was a computerised simulation for fire control, the behaviour of the simulated fire itself was complex enough to allow the users to input a large array of responses into the simulation and still have it model the impact on the simulated fire, while taking place in an abstract computerised task.

For our task, we are only interested in our participant's ability to determine between two separate pre-defined system states; threat detected or no threat detected. As such, the way the simulation combines the available data must be complex enough to create a challenge.

The simulation used in this research has been discussed at scientific advisory committees within the industry of security and defence and has been judged to be an acceptable standard for further psychological testing.

3 Experiment 1: Initial performance evaluation

3.1 Introduction

This experiment set out to evaluate the success of the simulation described in Chapter 2 as a decision-making paradigm under which anomaly handling can be examined. The task was one of hypothesis selection within a simulated intelligence-analysis environment. By creating low-frequency problems (anomalies) whose defining characteristics are similar to those of more frequent problems yet have distinct separate causal origins, it is possible to investigate an individual's ability to determine the difference between these problem types, and gain insight into the strategies used. Anomalous problems involved cases where the outcome contains cues that usually associate with the opposite position, for example, a no-threat system contains indicators usually indicative of a threat (complex environmental events that possess the weapon distribution of a conflict), or a threat system that contains indicators usually associated with no-threat (presence of environmental event). By creating such problems, it is possible to examine the degree of understanding the participant have of the problem-space and examine their underlying assumptions that represent their view of the relationships between objects in an environment. If a participant was drawing inferences set out in the rules and used a logical system of problem-solving, solution rates would be high. However, if heuristics are being used, then particular patterns of systematic error would indicate particular strategies.

Based upon the literature forming the recognition-primed decision making (RPDM: Klein et al., 2005) framework, we would expect four types of behaviour from participants in regards to strategies, each creating detectable signature in which errors are made:

- 1) The participant is using deductive reasoning to create inferences based upon the logical rules provided, this should result in no errors;
- 2) The participant is using a heuristic approach that dictates group membership of a

case based upon similarity to previous cases. This is a type of pattern-matching that is commonly associated with RPDM.

3) The participant is using a heuristic approach that dictates group membership of a case based upon the differences between a particular case, and those previously experienced. This is a type of change-detection that can be associated with the violation of expectation principle (Klein et al., 2005) described in Chapter 1.

4) The participants were using a combination of such strategies.

In particular, performance on the complex event problems can be used to determine which of these four outcomes is occurring. The complex problems were either complex conflict events or complex environmental events. The two heuristic strategies; change-detection (CD) and pattern-matching (PM) can achieve success on only one complex problem type. A CD strategy could achieve success on complex environmental events but would incur an error in the complex conflict events. A PM strategy would successfully identify complex conflict events but would systematically lead to failure in the complex environmental events.

The reason these strategies would result in the mentioned errors is due to which cues each heuristic approach would utilise. Pattern-matching (PM) is a rudimentary form of recognition-based decision making. This would materialise when a salient cue becomes indicative of the entire system state. Participants using this approach would associate the state of the simulation with the presence of a single salient cue. This cue could be either the presence of weapons, or the presence of an environmental event. For example, if participants associate a large rapid increase in weapons with a conflict, then they may regard the environmental data as superfluous and focus only on the weapons. It is possible to identify simple conflicts and simple environmental events using this approach only. However, for complex trials, this would achieve success on the complex conflict events – trials where there

is a conflict (identified via a large rapid increase in weapons) and an environmental event which may not be attended to. This PM strategy would, however, fail on the complex environmental events – as the weapons distributions profile is the same as a conflict, and has its genesis in multiple environmental events which can only be detected by monitoring the environmental data. If a participant is using only the single cue of the weapons data as a proxy indicator for the state of the system, then success would not be found on this problem type.

Similarly, change-detection (CD) also creates a detectable error pattern. Change-detection is a heuristic strategy based on the violation of expectation – when there is a cue in the environment which does not act like anticipated which may alert the participant that a change has occurred. For example, in simple conflict events, there is minimal change in the values corresponding to the environmental data, as there are no environmental events. As such, this environmental data remains relatively stable over a series of conflict trials. An example of a change –detection strategy would be when participants notice a change in the environmental data due to the onset of an environmental trial following a conflict trial, and thus reason a solution based on the new problem being different to the previous problem. If a participant is solving multiple simple conflict trials and becomes sensitive to the stability of the environmental data, then a disturbance in this environmental data may alert the participant that the problem requires an environmental event as the answer, and this would be correct for the simple environmental events. However, this creates issues in the identification of complex events. A participant using a CD strategy and solving problems based on the sudden presence or absence of environmental events would be expected to correctly solve complex environmental events (where there is an onset of an environmental event which is the cause of the change) and would lead to failure in the case of complex conflict events (where there is an environmental event that is not explanatory of the change in arms traffic). See Table 4 for

a comparison of behaviours between each strategy.

Table 4. Table showing how strategies are mapped to participants based on the responses to pairs of complex events.

Strategy	Complex Environmental	Complex Conflict
Inferential	Correct	Correct
Pattern-matching	Incorrect	Correct
Change-detection	Correct	Incorrect

The experiment contained two pairs of complex problems, and each participant's strategy was inferred from their performance on each pair of problems. Participants who correctly solved both complex problems in a pair were deemed to be using correct inferential reasoning. Participants who correctly solved both problems but demonstrated errors in other parts of the task, or incorrectly solved both problems, were categorised as using a possible mixture of the CD or PM strategies.

The other manipulation was mental set, to test for an Einstellung effect in anomaly handling. This is on the basis that familiarity encourages mental rigidity, and this rigidity may incur solution penalties when facing new types of problem. To integrate this into our experiment, the notion of high frequency normality was used to establish familiarity. In those we wished to establish a mental set, we provided a similar series of problems with similar outcomes. For this, the simple conflict problem type was chosen as it utilises the fewest cues in the simulation. It is then possible to test for mental set effects by subsequently providing participants with a problem type similar in appearance to the simple conflict but requires an alternative solution. For this, the complex environmental problem type was chosen, on the basis that the weapons distribution profiles of these two types of problems are identical. For

this reason, it is expected that, where a mental set has been established prior to the onset of the complex environmental problem, participants will identify similarities in weapons distribution and classify it similarly to those from the simple conflict problem type.

The anticipated behaviour of participants, as informed by the previous research described above, results in a number of hypotheses surrounding the use of the task, and the manipulation of mental set. Firstly, it is expected that participants will achieve higher solution rates on the simple problems than the complex problems. This is because complex problems require the interpretation of more cues from the simulated environment. The simple problems have a solution that favours the weighting of individual cues differently to those required for a successful solution. Secondly, those participants that experience an induced mental set by repetition of the same problem-type will suffer lower solution rates on a subsequent complex problem when it shares stimuli similarity with the problem-types that were used to establish the mental set. Thirdly, errors in the way that complex problems are answered will allow us to identify additional heuristic strategies used by participants.

3.2 Method

3.2.1 Participants

All participants were recruited from the University of Surrey. Undergraduate psychology students received a course credit for participation. A total of 24 individuals were tested (23 female, 1 male).

3.2.2 Design

A 2x2x2 mixed design was used for this study. *Set* was a between-subjects factor with two levels: *set* and *no set*. Participants were randomly allocated between these conditions. The difference between set and no set conditions was the order of presentation of stimuli. Those in the set condition received a series of simple conflict problems prior to the

complex environmental problems. In the no set condition, the complex environmental problems were preceded by mixed problem types.

One within-subjects factor was *Event*, with two levels; *environmental*, and *conflict*. Another within-subjects factor was *Complexity*, with two levels; *simple*, and *complex*. These factors combined to create 4 distinct problem types: *simple conflict event*, *simple environmental event*, *complex conflict event* and *complex environmental event*. All problems were completed by both groups.

3.2.3 Materials

The task was conducted at a computer using a digital version of the simulation explained in Chapter 2, the simulation was coded and programmed using bespoke experimental software, PsyScript (based on AppleScript). To summarise, participants were given a series of 18 trials to solve, concerning the security situation of a fictional country. In each trial participant had access to information on the distribution of weapons across geographical locations, along with the local environmental conditions at the time. The trials were presented within a simulated intelligence task. The simulation consisted of three visual elements; a map, a table containing the distribution of weapons, and a table containing environmental conditions. The map showed four zones alongside the supply routes by which weapons are known to be trafficked. A number of terrain features were also marked that intersected these supply routes; a mountain pass, a river, and a series of underground tunnels. The weapon distribution table showed the estimated number of weapons in each zone across five time-points. The table of environmental conditions gave information on three variables across those same five-time-points; rainfall, temperature, and recent seismic activity.

Each trial comprised of five time points across which the weapons distribution and environmental conditions were updated. Participants had to classify each trial into one of three categories: no event, conflict event, or environmental event. At the beginning of each

trial, all weapons were distributed equally between zones. As the simulation progressed, the supply of weapons could become asymmetrical, leading to large differences between zones in the number of weapons they possessed. The distribution of weapons during the simulation was determined by the presence of one of the three possible events.

The simulation was designed so that problems were explained exclusively by one of three possible hypotheses; scenarios where no significant arms traffic change had occurred, scenarios where arms traffic had changed due to environmental events, and scenarios where arms traffic changed due to conflict events. The correct response could be inferred from relationships between weapons supply, possible conflicts and environmental conditions, as follows:

- No changes in arms supply across zones over time indicated no presence of environmental or conflict events.
- A conflict is indicated by an accumulation of weapons in a single zone over time. Since weapons supply is finite, this leads to decrease in weapons supply to other zones.
- Changes in weapons distribution across zones might also indicate a change in environmental conditions. Each zone possesses terrain features along their supply routes (e.g., river, tunnels, mountain pass). High rainfall floods the rivers, recent seismic activity collapses tunnels, and large decreases in temperature freeze the mountain pass. When the supply route of a zone passes through a terrain feature affected by these conditions, that supply route becomes treacherous. Treacherous supply routes lower the number of weapons that can be delivered to those affected zones. When this occurs, surplus weapons are sent to regions with non-treacherous supply routes. For example, if Zone A requires crossing a river, and Zones B, C and D do not, then high rainfall would make the route to Zone A treacherous, leading to fewer weapons arriving in Zone A, which would then be redistributed evenly between Zones B, C and D. One or more zones

can be affected by closure of supply routes due to environmental conditions at any one time and which routes are treacherous can be determined by information provided in the environmental conditions table.

3.2.4 Procedure

Each participant was presented with written task instructions that were available throughout the experiment and included a set of rules that the simulation followed, including classification criteria for each problem-type. Participants then received on-screen instructions that introduced the visual elements of the task. Participants were given three practice trials that consisted of one of each possible scenario outcome (no event, simple environmental event, and simple conflict event). During the practice scenarios, participants were able to question the experimenter and were given the correct solution to ensure they understood the response options. After the three practice scenarios, participants completed a series of 18 problem trials. These trials comprised three no event problems, seven simple conflict-events, and four simple environment events. There were two *complex environmental events* where multiple environmental conditions interacted together to create a weapons distribution profile that appeared erroneously to indicate a conflict. The remaining two problems were *complex conflict events* that are conflict event scenarios that were made more difficult by the presence of a significant change of environmental conditions that occurred in irrelevant parts of the map. These four tasks required careful reasoning to discriminate between conflict and environmental explanations.

Participants in the *set* condition received a series of three conflict trials before being presented with the complex trials. In the *no set* condition, the exact same trials were used, but the order was changed to ensure that the problem types preceding the complex trials were mixed in nature between conflict and environmental trials. The order of presentation of the

trials can be found in Table 5 which illustrates the difference between set and no set conditions.

Table 5. *Order of presentation of stimuli between set and no set conditions.*

\boxed{X} = conflict events, $\boxed{-}$ = non-conflict events, $\boxed{0}$ = complex environmental event trials.

Condition	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Set	-	X	-	-	-	X	X	X	0	X	X	-	-	-	X	X	X	0
No set	-	X	-	X	X	-	X	-	0	X	X	-	X	X	-	X	-	0

As the simulation progressed, an on-screen prompt appeared to participants.

Participants were required to choose one of the three possible hypotheses they believed best described the current status of the system:

- i. No events occurred;
- ii. Change due to environmental events;
- iii. Change due to conflict events.

Once a decision had been captured for that particular problem, the next trial began.

This returned the tables to a state of equal distribution and represented the first time-point in a new scenario.

After instructions and familiarisation with the system, participants were instructed to complete the task as fast as possible while maintaining accuracy. Participants spent 30 minutes with the researcher from start to finish.

3.2.5 Coding Strategy

Performance on the complex problems can be used to determine strategy. Each of the two heuristic strategies; change-detection (CD) and pattern-matching (PM) can achieve success on only one type of complex problem type. The complex problems were either complex conflict events, or complex environmental events. A CD strategy could achieve success on a complex environmental event, but would incur an error in the complex conflict event. A PM strategy would reverse that, being able to successfully identify a complex

conflict event, but would systematically lead to failure in the complex environmental event. Participants who correctly solved both complex problems in a pair were deemed to be using inferential reasoning. Participants who correctly solved both problems but demonstrated errors in other parts of the task, or incorrectly solved both problems, could not systemically be associated with an exclusive strategy. Due to the nature of the strategies being coded, the strategy can only be determined across two responses to both complex problems. As such, the coded strategy is only computed by the performance across both complex problems and cannot be attributed to an individual attempt to solve a single problem.

3.3 Results

3.3.1 Analytical approach

Data gathered from the task were the hypothesis selected in response to each scenario, and the time taken to reach that selection. The responses were transformed into solution rates for each trial type to allow for comparisons between trial types despite unequal frequencies of each trial type. No-event trials were included only to ensure attention was paid to the task.

The variables of *event-type* and *complexity* combine to create 4 distinct types of problem described earlier. The majority of problems presented to participants in this experiment were simple problems. It is not expected that *Set* will influence simple problems. *Set* is expected to influence the complex environmental event trials as this is the critical problem that follows after the set has been established and shares similar features while requiring a different response. As such, it is not expected for *Set* to create an overall difference on solution rates. While each factor is examined independently, it is the interaction between these variables at this level that is the most interesting. As prior research shows that knowledge structure and heuristics are important to developing expertise, this will be kept as a two-tailed test as heuristic development may actually lead to the creation of new domain

specific strategies that could enhance problem-solving.

3.3.2 Solution rates

A repeated measures ANOVA was conducted on the proportion of correct responses across Complexity and Event-type, with Set as a between-subjects factor. No difference was found for proportion of total correct responses between Set ($M=0.82$, $SD=0.12$) and those in the No-set conditions ($M=0.83$, $SD=0.16$), $F(1,22)=0.19$, $p=.66$, $\eta^2=.009$, 95% CIs [0.77, 0.88] and [0.78, 0.89] respectively.

A main effect was found for Complexity, where Complex problems ($M=0.69$, $SD=0.37$) were solved less frequently than Simple problems ($M=0.89$, $SD=0.17$), $F(1,22)=14.02$, $p<0.001$, $\eta^2=.39$, 95% CIs [0.59,0.78],[0.84,0.94] respectively.

There was no significant main effect for Event type, $F(1,22)=0.67$, $p=0.79$.

A two-way interaction for Complexity by Event was significant, $F(1,22)=10.34$, $p<0.01$, $\eta^2=0.32$. Post-hoc tests show that complex conflict problems were solved less often ($M=0.58$, $SD=0.38$) than simple conflict problems ($M=0.97$, $SD=0.07$), whereas no difference was observed between complex environmental ($M=0.79$, $SD=0.36$) and simple environmental ($M=0.80$, $SD=0.26$) problems.

All remaining two-way interactions were not significant. These were Event by Set, $F(1,22)=0.60$, $p=0.45$, and Complexity by Set, $F(1,22)=3.68$, $p=0.06$.

A three-way interaction was found between Complexity, Event type and Set, $F(1,22)=6.98$, $p<0.05$, $\eta^2=.24$, where participants in the Set condition solved fewer complex environmental problems ($M=0.62$, $SD=0.14$) than participants in the No Set condition ($M=0.95$, $SD=0.43$). See Figure 3.1 for a chart of solution rates.

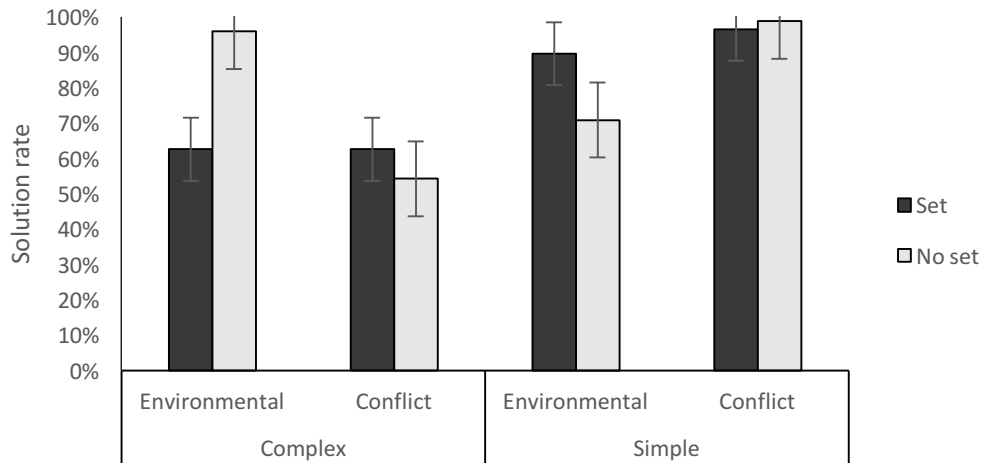


Figure 3.1. Mean proportion of correct responses for each problem type by set ($n=24$)

3.3.3 Solution Times

The solution time was the time taken (in seconds) to reach a decision about the classification of the trial.

An ANOVA was conducted on the solution times for Complexity and Event type by Set. Although time taken to solve complex problems was greater ($M=27.11s$, $SD=18.27s$) than the time taken to solve simple problems ($M=17.39s$, $SD=8.8s$), the main effect for Complexity only approached significance, $F(1,20)=3.78$, $p=0.06$, $\eta^2=.21$. There were no significant main effects for Set $F(1,20)=0.49$, $p=0.35$, or Event type, $F(1,20)=0.16$, $p=0.69$. The two-way interaction for Complexity by Event approached significance, $F(1,20)=3.71$, $p=0.056$. The two-way interactions of Complexity by Set, $F(1,20)=0.10$, $p=0.75$, and Set by Event, $F(1,20)=1.19$, $p=0.27$ were also not significant. The three-way interaction for Complexity by Event by Set was also not significant, $F(1,20)=0.41$, $p=0.52$. See Figure 3.2 for correct solution times for each specific problem type by set.

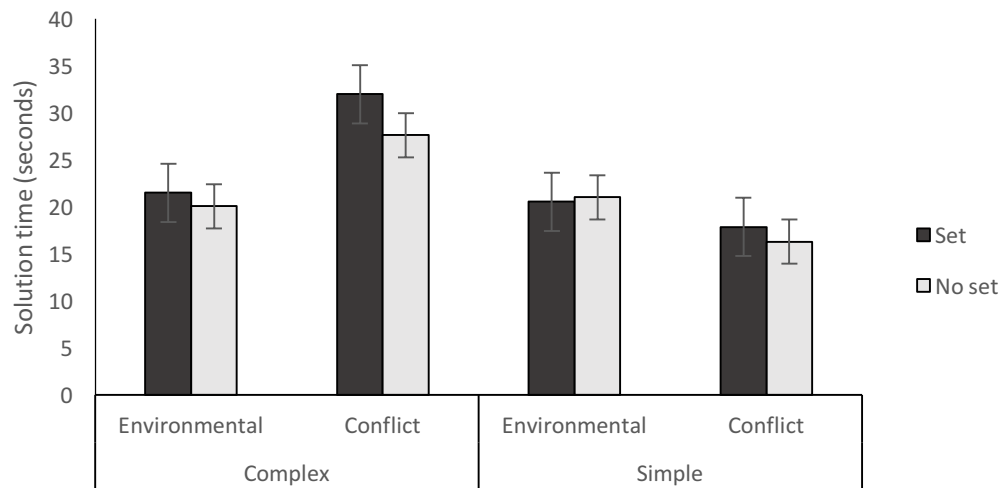


Figure 3.2. Correct solution time (s) for each specific type of problem by set ($n=24$).

3.3.4 Strategy use

The problem pairs were coded according to the four strategies outlined in 3.2.5. Of the 48 problem pairs coded, 9 could not be attributed to an exclusive strategy, and only the remaining 37 were included. Figure 3.3 shows the frequency of use of each strategy between each group. The pattern-matching (PM) strategy occurred more frequently in the Set group than the No Set group. The change-detection (CD) strategy occurred more frequently in the No Set group. Both groups had equivalent numbers of participants using both inferential strategies. A Chi Square showed that these differences in frequency of strategy use approach, but were ultimately not significant, $\chi^2(2, N=37) = 5.95, p=0.051$.

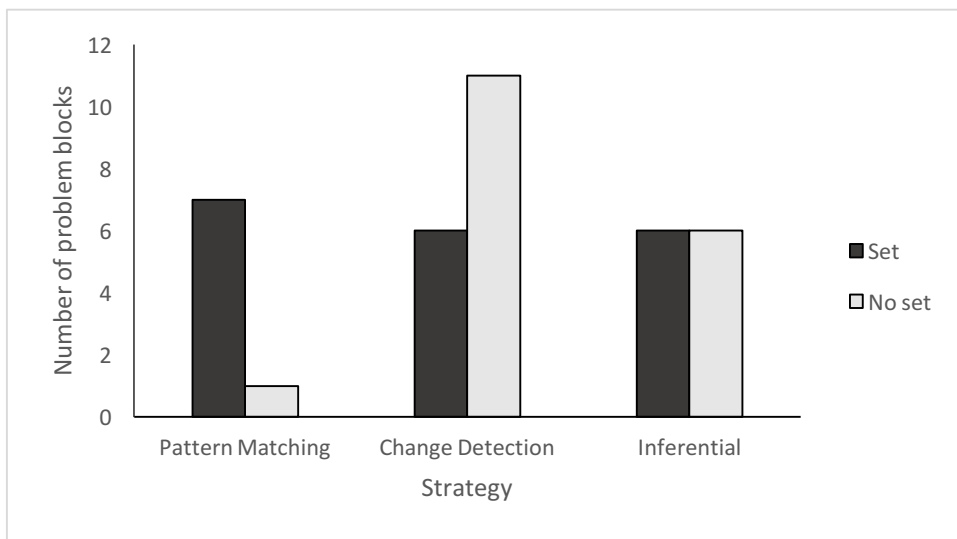


Figure 3.3. Frequency of each strategy use across all problem-pairs between the set and no set groups.

The above figure examines the aggregate of all problem-pairs. It is also possible to examine the frequency of strategy use for each problem-pair separately, so that change in strategy over time can be examined. The first block of problem-pairs, Block 1, was the first occurrence of a problem-pair within the task. This occurred within the first 9 trials. Block 2 was the second occurrence of the problem-pair that occurred between trial 9-18.

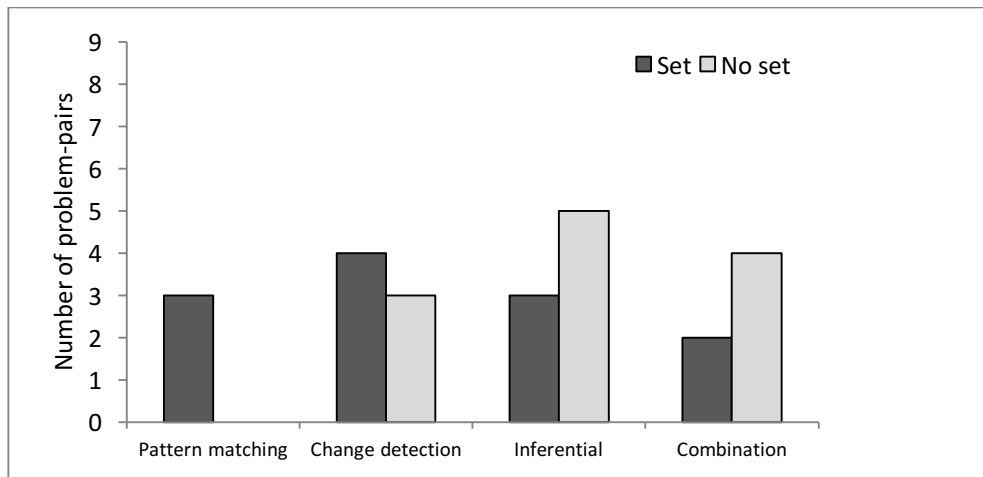


Figure 3.4. Frequency of strategy use as identified in Block 1 problem-pairs.

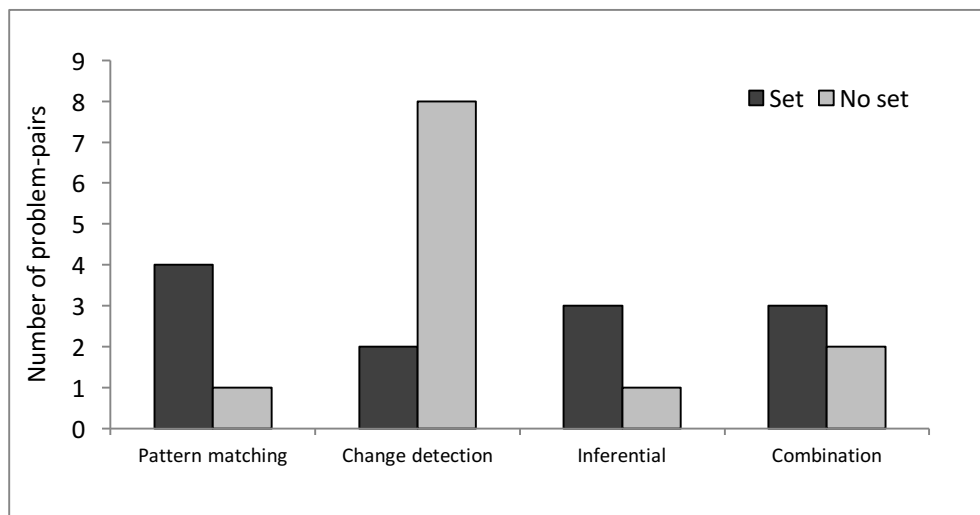


Figure 3.5. Frequency of strategy use as identified in Block 2 problem-pairs.

Figures 3.4 and 3.5 show the frequency of use of each strategy between set and no set groups at each of the two points in time. Strategy use appears to change over time. Most notably, the no-set groups move away from inferential and combined strategies towards a CD strategy. The set group appears to maintain their use of inferential and combined strategies, with a small move from CD towards PM strategies.

3.3.5 Strategy and Response times

An ANOVA was conducted to re-analyse response time (in seconds) for each strategy. To identify these cases, the response times for trials in the problem-pairs were averaged across responses from both response times to both problems within the problem-pair, and re-analysed based on the identified strategy. A significant effect was found of Strategy, with inferential strategies ($M=34.34s$, $SD=24.67s$) taking longer than those using pattern-matching ($M=22.07s$, $SD=14.91s$) and change-detection ($M=23.48s$, $SD=18.46s$) strategies, $F(3,88) = 2.77$, $p = 0.01$, $\eta^2=0.18$. A significant interaction was also found, $F(3,88) = 3.25$, $p = 0.025$, $\eta^2=0.10$. Post hoc tests show that for the pattern-matching strategy, the set group took longer than the no set group. For the inferential strategy, the set group were faster than the no set group. All response times for strategy and group type are shown in Figure 3.6. While it is notable that the pattern-matching strategy found in the no set group had an extremely low average response time, it should be pointed out that this PM/No set strategy was only identified in one problem-pair and so represents a single data point from a single participant.

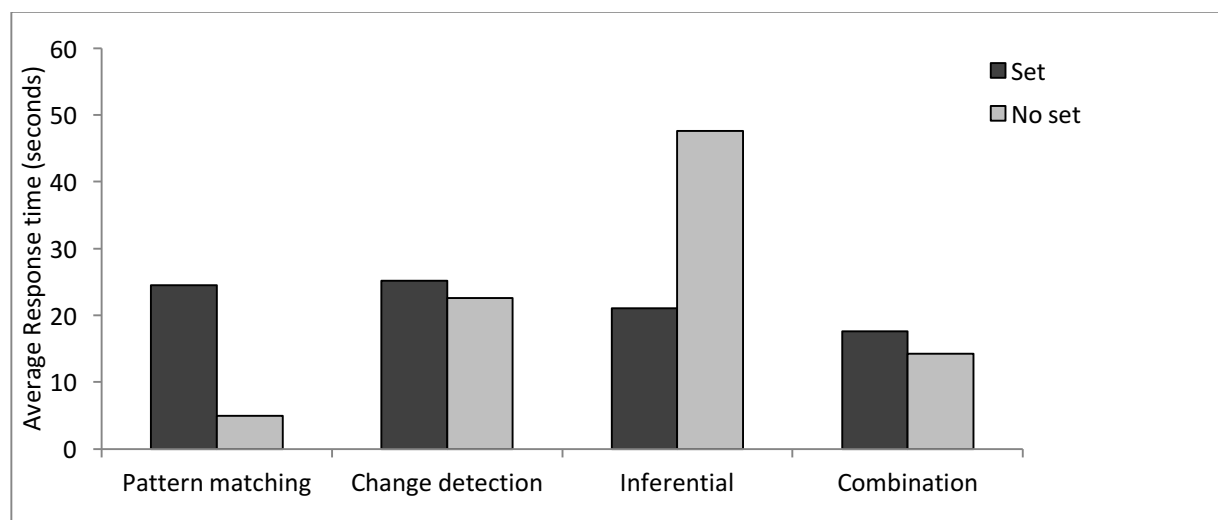


Figure 3.6. Mean response times (seconds) for average problem-pair responses for strategy by set.

3.4 Discussion

In this study, we examined the solution rates and times in a security assessment task consisting of sets of simple and complex trials. The rationale for including the complex problems was that they are representative of anomalies. The complex environmental event presents a combination of environmental factors that combine to create a weapon distribution profile unlike simple environmental events, but similar to a simple conflict event. Incorrectly classifying them as the latter is a type of false positive, a scenario that has an appearance of a threat, but actually is harmless. This is akin to the type of error that drone operators make when striking a wedding instead of a group of militants. Conversely, complex conflict event problems could only be explained by a conflict event, but the presence of an environmental factor made them appear similar to simple environmental problems. Incorrectly classifying them as the latter is a type of false negative. In order to solve these problem types successfully, participants have to reason according to the rules and overcome heuristic responses based on cue saliency.

The study also examined the phenomenon of mental set, in which the presentation order of problems was manipulated prior to the onset of complex event scenarios. Participants in the set condition received a series of three conflict event trials prior to a complex environmental trial (representing an anomaly). As a result, they solved fewer trials than participants in the no-set condition. The results indicate that the similarity in cue characteristics creates poorer performance for those in the set condition. Participants' previous success at categorising the conflict problems based on their weapon distribution profiles may lead them to put more weight on the diagnostic relevance of the distribution profile, and subsequently rely on a strategy of *pattern-matching* in which the weapons distribution profile is examined and determine to either match or not match the typical profile

of that problem type. This would explain the high solution rates of other problem types, and explains the failure to identify the complex environmental problem. Participants in the set group who then attempted to solve the complex anomaly problem using a pattern-matching approach would look to see if any new problems had a weapon distribution pattern consistent with a simple conflict event and could have paid less attention to the environmental conditions, which had not been diagnostically relevant up until that point. Participants in the no-set group would have been more likely to attempt to use all of the information available to them, allowing them to detect changes in environmental conditions and subsequently classify the problem correctly.

The idea that some participants may be using a pattern-matching strategy does not explain the high solution rates for the complex environmental event problems for participants in the no set group. To account for these results, another strategy must be introduced, one of *change-detection*. During the critical false positive trial achieved through the use of a complex environmental event, participants may recognise the distinctive distribution profile of a conflict; however, they also are capable of detecting change in the environmental information. If participants recognise the weapons distribution as similar to a simple conflict event, but detect a change in environmental conditions that the previous simple problem type does not have, then they could reason that the situation must be different from the simple explanation and choose the remaining alternative.

The other problem type of interest is the false negative given by complex conflict events. This problem had the lowest mean solution rates across both groups. For this problem type, a scenario was presented in which the weapons distribution profile indicated a conflict event. However, an environmental indicator was also present in an irrelevant region to the conflict. In effect, there is a conflict event occurring, but also an irrelevant environmental effect that acted as noise. In order to achieve success on this problem, participants would

have to recognise the weapons distribution as being created due to a conflict event only, and reason that the final distribution could not be caused by the environmental event present.

Using a change-detection strategy, participants would recognise the distinctive conflict-like weapons distribution, but that they also perceive the environmental effect that could lead them to classify the scenario as one being different to a conflict event. This approach worked for the solution of the complex environmental event, but here would be incorrectly applied to this problem and lead to a failure to identify the conflict that is driving distribution in favour of merely noticing a perceptual difference between problems.

Conversely, those using a pattern-matching approach may prioritise the diagnostic relevance of the weapons distribution and ignore the environmental information. For complex environmental events, this strategy leads to an error in failing to identify the true cause of a situation. However, for complex conflict events, a pattern-matching strategy could provide a correct solution, since participants pay more attention to the weapon distribution. For the complex conflict event, an account of pattern-matching and change-detection would lead to those using a pattern-matching strategy to correctly solve more of this problem type than those using a change-detection strategy. It was predicted that those in the set condition would correctly identify this scenario more often than those in the no set condition. While our results support the direction of this prediction, the difference between set and no set was not significant for this problem type.

The use of pattern-matching and change-detection heuristics appears to explain errors with complex problems, yet the majority of scenarios faced by participants are simple problems (simple conflict event, simple environmental event, no event) that can be solved using these heuristics. As such, they represent an efficient method of decision-making. In this study, there were 14 simple problems, and 4 complex problems, if each strategy is only failing on 2 of the complex problems, then a change-detection or pattern-matching approach

will be successful on 16 out of 18 trials. The simulation works around a core system of logical statements that dictate the distribution of weapons based upon a number of factors. Both pattern-matching and change-detection strategies ignore this logic in favour of interpreting the outcome of the system based upon a suboptimal strategy developed during the solution of simple problems.

However, results from solution time data suggest that participants are taking longer to solve complex problems. If participants were adhering to a single strategy approach as discussed, then we would expect participants to respond in the same timeframe for complex problems as for simple problems, as decisions would be based upon perceptual information from the simulation environment rather than reasoning. However, the increased solution time for complex problems suggests that participants are aware of the conflicting information being presented to them. Perhaps an awareness of the entire environment is present and participants attempt to reason the inconsistencies observed in complex problems, but when this becomes difficult participants fall back onto their strategies that have worked for simple problems leading to error.

This unexpected finding aligns with two previous areas of psychological research. Fasolo, McClelland and Lange (2005) found that conflicting evidence increases the perceived difficulty of a task. Bettman, Johnson, Luce and Payne (1993) found that participants are willing to trade accuracy for effort. It may be that by detecting a conflict in the information streams, participants then realise the task is more difficult and trade off accuracy for effort by employing the suboptimal heuristics. This would explain the lower solution rates on complex trials whilst still having higher solution times.

Another area that could have impacted is that of perceptual error, where participants simply failed to notice a relevant cue in the task. Simple environmental problems had weapons distributions altered due to a presence of environmental events. However, the level

of weapons in each zone never reaches the heightened levels seen for conflict events. If environmental events made supply to one zone treacherous, this zone would have its supply redistributed along remaining safe routes. Similarly, if the supply routes to two zones were treacherous, the combined supply would be redistributed along remaining safe routes. Thus, a small number of weapons may be shared by three zones, or a larger number of weapons may be shared by two zones. The number of affected supply routes may affect a change-detection heuristic because two-zone scenarios show a more marked change in weapons distribution than one-zone scenarios, which mean that the level of change is too small for participants to detect for some trials.

Change-detection may also be influenced by the variables used to represent environmental information. The scales used for rainfall and temperature are large and changes within them are easy to spot, but seismic activity was represented using the Richter scale, a base-10 logarithmic scale where changes are smaller. Changes in rainfall from 7mm to 70mm or in temperature from 16C to -16C are noticeable, but a change in Richter from 5 to 6 may not be clear. This may explain why solution rates for simple environmental events are lower than those of simple conflict or no-event problems. The Richter scale represented recent seismic activity. Scenarios involving changes caused by seismic activity had one of the lowest solution rates, even lower than some of the complex problems. Participants using a change-detection approach failed to detect subtle change in this indicator and did not consider it a relevant cue. This would explain a number of participants who incorrectly categorised this scenario as a no event scenario. Where participants focussed on the weapons distribution information they had to account for a change in weapons when there appeared no environmental event, which could lead them to classify the scenario as a conflict. This suggests that change-detection is a generic strategy that was used for simple as well as complex problems.

More participants in the set group used a pattern-matching approach, while those in the no-set group tended to use change-detection. The set group presumably weighted the information provided by the weapons distribution profile higher than the information provided by environmental conditions. Weapons distribution alone is capable of solving the set-congruent problems and this strategy did not change when the nature of the problems did. This appears to follow the pattern of choosing a sub-optimal solution strategy due to familiarity and repetitiveness when a more appropriate alternative solution existed, as seen in classical mental set (Einstellung effect) literature outlined in Chapter 1. It is likely that participants developed dependency for using the weapons information as the only cue, and learnt to discriminate the different simple event trials using a single cue. Examination of the frequency of use of strategy was also examined across the two problem blocks. For the set group, strategy change was minor, with a slight convergence towards pattern-matching. However, for the no set group, a rapid and strong change was noticed with participants more likely to be using a change-detection strategy on the second block than the first. It may be that the no-set group used inferential reasoning more frequently in block one because of unfamiliarity with the system. After familiarisation, participants can begin to identify heuristics for making their decisions and shift to a less cognitively expensive strategy. For the set group, mental set may restrict the search for alternative explanations since identifying a recurring pattern seems to work effectively. This may be why set showed continued reliance on weapons distribution in the development of pattern-matching strategies which failed on the complex environmental tasks.

One of the more prominent issues with this experiment is that it showed a detriment in performance over time. Some participants started off using inferential strategies (and thus were attaining higher solution rates in the first part of the task) before switching to heuristic strategies (which result in lower solution rates later in the task). An extension of this is that

participants were using heuristics from the start and made critical errors on complex trials throughout. We also found that establishing a mental set lowered solution rates on complex problems that share surface-level similarities. This is an issue because once participants begin to make errors, there is no way for participants to know that what they are doing is not successful. As the goal of this research is to prevent these mistakes, participants need to know when they are making them. This concept was the design focus of the next experiment.

A second issue with this research was the small size of trials completed, at only 18 trials, there was only two instances of each of the complex event trials. As solution rates were calculated proportionally to compensate for unequal frequencies, this may impact the complex problems more as a single error on the first complex trial has a bigger impact on solution rates for those complex events. However, this is difficult to control for, as by vastly increasing the number of complex trials takes away their ability to be an ‘anomaly’ and will simply result in them being recognised as just another type of problem.

4 Experiment 2 – Feedback

4.1 Introduction

Experiment 1 found differences in problem-solving strategy between the set and no set groups. However, there was little improvement over time; instead performance actually decreased over time as participants developed heuristics they believed suitable for problem solving that led to errors. This can be seen in the no set group switching from an inferential strategy to a heuristic strategy of change-detection as the simulation progressed within a single session. The low solution times of the no set group using pattern-matching strategies suggest that very little information was being attended to and further suggests reliance upon inaccurate heuristics.

It is possible that performance can become worse over time as there was no way for participants to test how well they were doing as no feedback was administered until after the experiment. Participants solving problems during the experiment would likely become confident in their reasoning and use that same reasoning again to tackle future problems, as we have seen in the research for cognitive rigidity in Chapter 1. However, participants were never informed that their reasoning may be inaccurate, or when their current strategies were unsuccessful. This can be eliminated if feedback was introduced after each problem, as it would then signal to participants that the strategies they are using are not working.

Feedback leads to further testable hypotheses. Firstly, if participants are using heuristic strategies, then the introduction of feedback should increase solution times, as participants may slow down to analyse responses in more detail after they have been informed their chosen solution was incorrect. By introducing feedback, we may also get a better test of the true effects of mental set, as feedback provides an opportunity for corrective behaviour in regards to incorrect solution strategies. Secondly, introducing feedback should allow participants to achieve higher solution rates across all problem types, as fixation on

heuristics may be lessened as participants are informed when they are unsuccessful. This will not only be further evidence that heuristics were used in Experiment 1 but also show how flexible participants can be to changing strategies when they are discovered to be inadequate. However, due to the nature of the problems, it is still expected that complex environmental problems will be solved the least due to the complexity of the strategy required for solution.

The number of trials was increased to overcome the disproportionate impact of errors for complex trials. However, while this was changed from Experiment 1 to increase the reliability of solution rates, it does eliminate the ability to attribute the pattern-matching and change-detection strategies. The attribution of these strategies was only possible by comparing responses over problem-pairs. With the increase in the number of complex problems in this experiment, there is no longer a single problem pair to be coded.

4.2 Method

4.2.1 Participants

All participants were recruited online through social media advertisement. The web address was circulated online among academic interest groups in order to attract suitable participants. Participants were entered into a random prize draw upon completion of the experiment. An additional prize was awarded based upon performance to incentivise taking the experiment seriously. A total of 56 individuals completed the experiment (34 male, 22 females). Age of participants ranged from 18 to 53 ($M = 23.5$, $SD = 7.1$). Although the geographical location of participants was predominantly from the United Kingdom ($n=40$), other countries participating included the USA ($n=6$), Canada ($n=2$), Germany ($n=2$), Hungary ($n=1$), Kenya ($n=1$), Norway ($n=1$), Poland ($n=1$) and Sweden ($n=1$). Participants had to click a confirmation button that they were fluent in English. A further 31 participants began the experiment but did not finish, these participants were considered to have

withdrawn and were excluded from analysis.

4.2.2 Design

A 2x2x2 mixed design was used for this study, identical to the previous experiment. The between-subjects factor of *Set* was retained (*set*, *no set*). The within-subjects factor of *Event* was retained consisting of two levels; *environmental*, and *conflict*. Another within-subjects factor was *Complexity* that also consisted of two levels; *simple trials*, and *complex trials*. These factors combined to create 4 distinct problem-types that were identical to those in Experiment 1. The key difference between this experiment and Experiment 1 was the inclusion of immediate feedback after every trial.

4.2.3 Materials

The paradigm used in this experiment was identical to Experiment 1 except now it was hosted on the internet using a web-based platform. For this experiment, an increased number of trials were used. A total of 54 trials were created with 24 simple conflict events, 6 simple environmental events, 12 complex conflict events, 6 complex environmental events, and 6 no-event problems.

The experiment consisted of 6 problem-blocks, each containing a number of problems from various problem types, but always culminating with a complex environmental event. In the set conditions, this problem was directly preceded by a series of 4 simple conflict problems. In the no-set condition, this problem was preceded by a mixture of problem types.

4.2.4 Procedure

The experiment was hosted online using a researcher-owned web address (<http://www.anomalyhandling.com>). Participants visiting the web address were welcomed,

presented with an information sheet and consent form. Participants were informed that participation was voluntary and that they could exit the study at any time by closing the web browser. Continuing with the experiment yielded a comprehensive set of instructions that informed the participant about the system they were about to use. Rules were explained, examples were given, and participants were allowed to explore this until they felt ready to begin (these instructions can be found in Appendix 1). Once participants were ready they could begin the experiment. The procedure was identical to Experiment 1, except that after each decision is submitted, the participant received feedback informing them of the correct response to each problem. This feedback was shown before they proceed to the following problem, and was present after every trial.

4.3 Results

The data comprise the categorisation of the event type for each trial, and the time taken to make that decision. A problem was considered correct if the categorisation was the same as the target classification. The responses were transformed into solution rates for each trial type to allow for comparisons between trial types despite unequal frequencies of each trial type.

4.3.1 Solution rates

Figure 4.1 shows the mean solution rates for each condition. An ANOVA was conducted on the proportion of correct response across Complexity, Event and Set. No difference was found in the solution rates between the set condition ($M=0.94$, $SD=0.03$) and the no set condition ($M=0.93$, $SD=0.07$), $F(1,53)=0.14$, $p=0.71$, $\eta^2=.003$, 95% CIs [0.91,0.98] and [0.90,0.97]. A main effect of Event was found, $F(1,53)=5.18$, $p<0.05$, $\eta^2=.09$. Environmental problems were solved significantly less often ($M=0.91$, $SD=0.15$) than conflict problems ($M=0.96$, $SD=0.06$), 95% CIs [0.87,0.95] and [0.94,0.98]. A main effect

was also found for Complexity, with complex trials ($M= 0.89, SD=0.14$) solved significantly less often than simple trials ($M=0.99, SD=0.05$), $F(1,53)=36.62, p<0.001, \eta^2=.41$, 95% CIs $[0.85,0.93]$ and $[0.97,1.0]$.

There were no significant interactions, though the Complexity by Event approached significance, $F(1,53)=3.69, p=0.06$ and trended in the predicted direction, with complex environmental trials having lower solution rates than complex conflict trials. The other non-significant two-way interactions were Complexity by Set, $F(1,53)=0.16, p=0.69$, and Set by Event $F(1,53)=0.42, p=0.84$. The three-way interaction between Set, Complexity and Event was also not significant, $F(1,53)=0.22, p=0.64$.

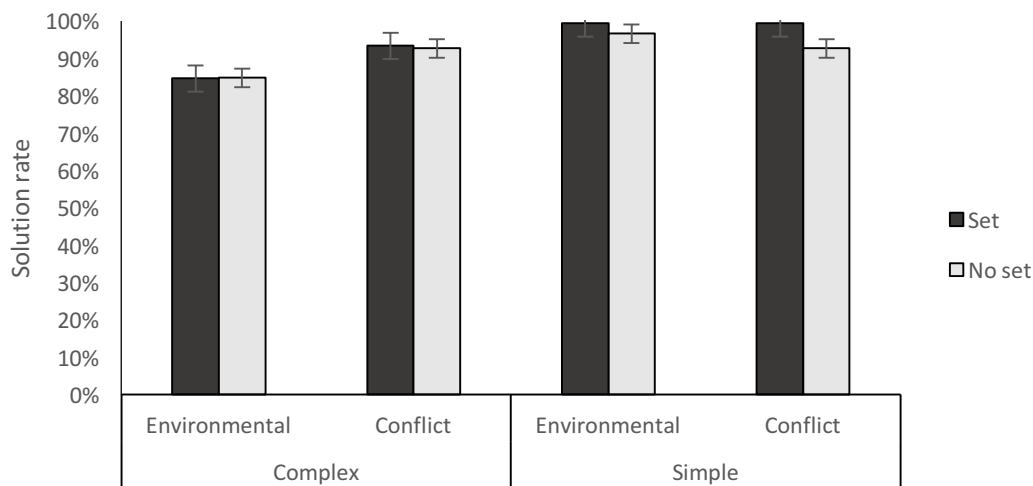


Figure 4.1. Mean proportion of solution rates for event-type and complexity by set ($n=56$).

4.3.2 Solution times

Figure 4.2 shows the mean solution times (in seconds) for each condition. An ANOVA was conducted on the mean solution times across all independent variables. The effect of Set was not significant, ($M_{set} = 33.94s, SD=15.58s$; $M_{no\ set} = 33.01s, SD=9.15s$), $F(1,53)=0.02, p=0.88$. A main effect was found for Complexity, with simple trials ($M=28.96s, SD=17.48s$) being solved faster than complex trials ($M=38.00s, SD=28.41s$),

$F(1,53)=15.38, p<0.001, \eta^2=0.23$, 95% CIs [24.07s,33.84s] and [30.05s, 45.95s]. Event was not significant ($M_{conflict} = 35.23s$; $M_{environmental} =31.72s$), $F(1,53)=1.09, p=0.30$.

A significant interaction was found for Complexity by Event, $F(1,53)=4.34, p<0.05, \eta^2=.079$, where complex environmental trials ($M=43.07s, SD=49.61s$) took significantly longer to solve than complex conflict trials ($M=32.96s, SD=14.26s$). Other interactions were not significant. These were Complexity by Set, $F(1,53)=0.38, p=0.54$, Event by Set, $F(1,53)=0.41, p=0.84$, and the three-way interaction of Complexity by Set by Event, $F(1,53)=1.21, p=0.27$.

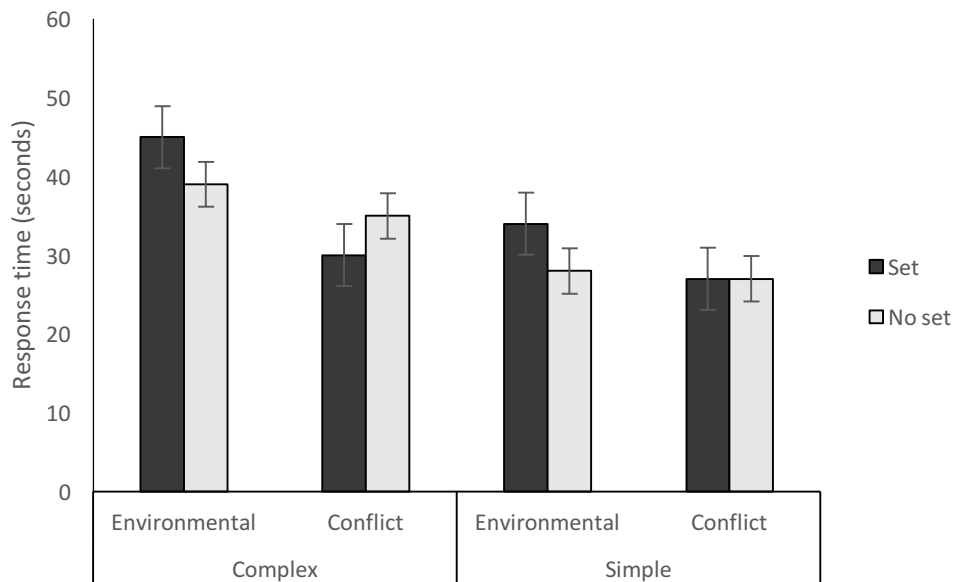


Figure 4.2. Mean solution time (in seconds) across complexity and event by set ($n=56$).

4.3.3 Learning over trials

This experiment comprised of six blocks of problems that each culminated with a complex environmental event problem. Solution rates on this problem were captured for each problem-block, and an ANOVA was conducted with Problem block and Set as factors. Mean solution rates on block 1 ($M=0.77$) were lower than block 2 ($M=0.79$), block 3 ($M=0.81$),

block 4 ($M=0.90$), block 5 ($M=0.92$), and block 6 ($M=0.87$) although these differences narrowly missed significance, $F(5,250)=2.17$, $p=0.058$, $\eta^2=.042$. There was no significant effect of Set, $F(5,250)=0.1$, $p=0.76$, $\eta^2=.002$. There was no significant interaction for Set by Problem-block, $F(5,250)=1.03$, $p=0.40$. See Figure 4.3 for the solution rates of complex environmental event problems across problem blocks for set.

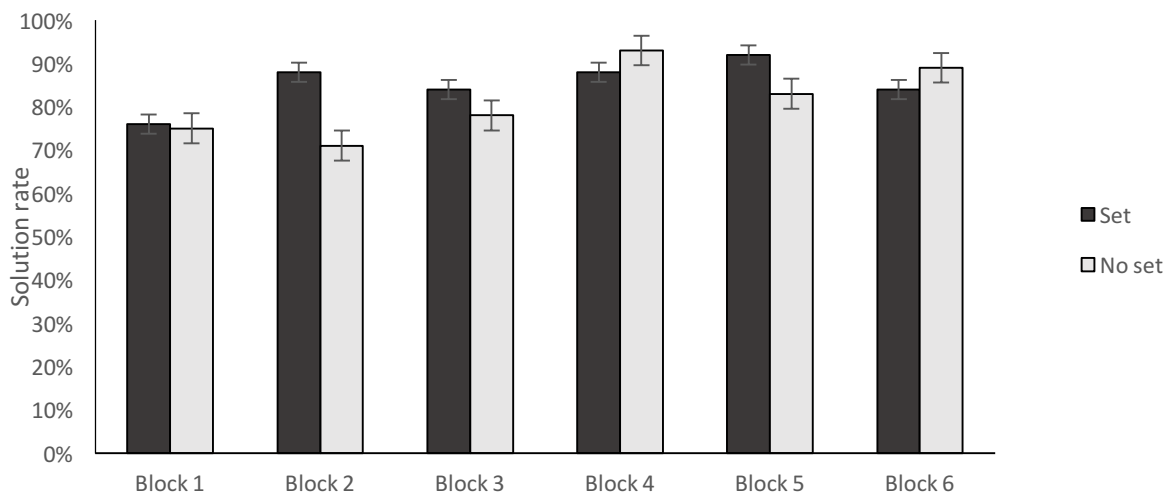


Figure 4.3. Mean solution rates for complex environmental event problem across problem-block between set and no set ($n=56$).

4.4 Discussion

In Experiment 1, the introduction of a mental set (established with simple trials) inhibited successful solution of a complex trial when the set problem and complex problem shared surface similarities. The present study has shown that this deleterious effect of mental set can be negated by the introduction of immediate feedback. In Experiment 1, it was suggested that participants, instead of utilising the causal rules of the system to make decisions inferentially, instead relied upon a series of recognition-based heuristics. By introducing feedback in the current study, participants discovered these heuristics were not

appropriate and either adjusted their strategies to include drawing inferences from the information based upon the system logic or refined their recognition-based heuristics to be more discriminative between problem types.

Introducing feedback had other consequences besides minimising the effect of set. Solution rates across all problem types were higher relative to Experiment 1, showing that feedback was not just alerting participants to potentially inappropriate strategies towards particular problems, but is also improving access to more optimal strategies across all problem types.

However, while feedback improved performance on all problem types and eliminated the effect of mental set, there remained an effect of problem complexity. Complex events were still the most difficult to identify, and complex environmental events (false-positives, where a suspicious indicator is present that is capable of being innocently explained) were the most difficult to solve. If feedback truly was improving strategy selection and participants were shifting from recognition heuristics to drawing inferences, then we would not expect this difference in performance, as participants reasoning by the system logic should attain equal solution rates in all trial types.

Rather than switching to inferential strategies and conducting causal reasoning around the system logic, it could be that participants were instead refining their recognition strategies by re-evaluating the importance of each cue in the environment and improving upon their ability to discern perceptual differences within the data presented to them. Potential evidence for a refined recognition heuristic comes from performance over time across the complex environmental event trial type, the most difficult trial type in the study. The average response time improved as exposure to the complex environmental event was increased. However, while this explanation sees the mean moving in the desired and expected direction, these differences between problem blocks narrowly missed statistical significance ($p=0.058$).

One potential weakness shown in this study is that by increasing the number of trials from 18 to 54, there were now a much larger number of trials of each type. This may explain why we have seen increased solution rates compared with Experiment 1. In Experiment 1, complex trials acted as anomalies, were low frequency events (compared to the simple trials), and as such, participants were less exposed to them and less experienced at drawing the required inferences for a correct solution. Now the number of trials has increased, it could be that the complex trials, while still appearing less frequently than the simple trials, are no longer rare enough to be true anomalies, and allowed participants, with the aid of feedback, to simply recognise them as another type of distinct trial type. However, the fact that complex problems were still solved less often than simple problems shows that there is a difference in the structure of the problems that participants were unable to grasp successfully.

By increasing the number of problems given during the experiment, it also eliminated the ability to attribute the pattern-matching and change-detection heuristic strategies. The attribution of these strategies was only possible by comparing patterns of responses over complex problem-pairs. With the increase in the number of complex problems in this experiment, there was no longer a single problem pair to be coded. While this meant that strategy was not able to be mapped as directly as before, the experiment benefitted from having more complex problems solved which allowed solution rates to vary over a larger number of problems.

While feedback appeared to remove the impact of mental set, it could be that the lower performance on complex problems comes down to how participants were weighting the cues within the paradigm. As discussed in the introduction of this thesis, acquisition of cues from the environment are critical to successful decision-making (Dawes & Corrigan, 1974) and it may be that by increasing the number of trials in the experiment, and exposing participants to a greater number of trials that can be solved using a single cue, that

participants still develop simple heuristics that use minimal cognitive effort by focusing on as few cues as possible. In order to try and promote a true picture of what this ability to solve complex problems looks like, Experiment 3 was conducted in which the cost of errors was manipulated to elicit more effort from the participants in order to explore the presence of an accuracy/effort trade-off, and its importance in the development of heuristics.

5 Experiment 3 - Cost

5.1 Introduction

Experiment 1 found that establishing a mental set can have negative consequences in solving anomalous problems. Experiment 2 showed that providing immediate feedback to participants minimised the negative effect of mental set. However, while achieving solution rate parity between set and no set conditions, feedback still failed to produce equivalently high solution rates between simple and complex problems. Complex problems were still not being solved as often as simple ones, even when the failed strategies used by participants were highlighted.

The primary reason for investigating anomaly handling was to provide an account of potentially high-risk decision making for low frequency events. In such cases, it is imperative that every problem is handled successfully and not just the simpler ones. This thesis has also directed its experimental work towards the domain of intelligence in the security and defence sector. In this industry it is also sometimes not possible to be able to provide feedback to intelligence analysts until a long time after the event, if at all. For example, if satellite imagery is analysed and suspicious activity is noticed, an analyst may report this for tasking. The outcome of that tasking may never be revealed to the analyst, due to compartmentalisation of process in the defence industry, or because follow-up observations are not conducted for a long time. As such, immediate feedback is not always available, and so other ways to minimise mental set should be sought and to improve the decision-making process so that all problems are able to be handled to the best capability of the decision-maker rather than relying on heuristics that only work in standard cases.

One such potential method is the introduction of cost. Anomaly handling is often associated with high-risk decision making. The penalty for failing to make a correct decision

can be catastrophic.

Experiment 3 explored the impact of cost of errors on the performance of participants using the anomaly-handling intelligence paradigm. This was measured by the introduction of a *cost* versus *no cost* group. It was hypothesised that those in the *cost* group will take longer to solve problems than those in the *no cost* group, and that there will be an increase in solution rates for the *cost* group over the *no cost* group.

The cost that was added to the paradigm was in the form of a financial penalty for errors made during the problem solving task. The *cost* group lost money for each incorrect decision made, while the *no cost* group suffered no penalty for error. Participants in the *cost* group began the experiment with a set amount of money (£5.40, representing 10p per trial), and subsequently lost 20p per incorrect response given. Participants in the *no cost* group received the full amount of their reward for participation and suffered no negative outcomes for incorrect solutions.

5.2 Method

5.2.1 Participants

A total of 70 undergraduate psychology students recruited from the University of Sussex received course credits for participation (50 females, 20 males). Ages ranged from 18-41 ($M=19.5$, $SD=3.07$).

5.2.2 Design

A 2x2x2x2 mixed design was used for this experiment. There were two between-subjects factors and two within-subjects factors. The between-subject factor of *Set* was retained (*set*, *no set*). A second between-subject factor was *Cost*, consisting of two levels (*cost*, *no cost*). Combining these factors creates four separate subgroups and participants were

randomly assigned to each: *set x cost* (18 participants), *set x no cost* (19 participants), *no set x cost* (18 participants) and *no set x no cost* (15 participants). Within-subject factors of *Complexity* (*complex trials*, *simple trials*) and *Event-type* (*conflict event*, *environmental event*) were retained. As with Experiments 1 and 2, these created four distinct trial-types; *simple conflict event*, *simple environmental event*, *complex conflict event*, and *complex environmental event*. As in previous studies, *no event* trials were incorporated into the task as a control to check for participant awareness and attention.

5.2.3 Materials

The task used in this experiment was identical to the task used in Experiment 2. The paradigm is a classification task where participants must evaluate each trial and categorise it based upon event-type (conflict or environmental). Instructions were included within the system that detailed how to complete the task and explained the rules of the system. A total of 54 trials were completed by all participants. These 54 trials consisted of 24 *simple conflict* trials, 6 *simple environmental* trials, 12 *complex conflict* trials, 6 *complex environmental* trials and 6 *no event* trials.

Trials were arranged into six equal blocks. All blocks culminated in a *complex environmental* trial. For the *set* group, this complex environmental trial was preceded by a series of four *simple conflict* trials, whereas in the *no set* group it was preceded by a variety of different trial types. Each block contained the same trials, the order of presentation being the only difference for *set* and *no set* groups.

5.2.4 Procedure

Each participant was tested alone. Participants were given information sheets explaining the experiment and were asked for their informed consent. Once consent had been obtained, participants were guided to use the instruction material built within the system to

learn the elements of the simulation display and the governing rules of the paradigm. Once the training had been completed, participants were given time to ask questions or clarifications to the researcher. Participants then began the experiment and were presented with the 54 trials. Participants responded to each trial by manually selecting a written statement on the screen that matched their evaluation of the trial.

For both groups, each participant was told they had a starting pool of £5.40 (representing 10p per trial). Those in the cost group were told that they would lose 20p per incorrect solution, and so a solution rate of greater than 50% was required for them to earn anything. Participants in the no cost group were told that they would receive the full amount of £5.40 regardless of performance.

Feedback was shown after each trial so that participants in the cost group knew when an error had been made and a cost had been incurred.

5.3 Results

The data comprise the categorisation decision for each trial and the time taken to reach that decision. A problem was considered correct if the participants' categorisation was the same as the trial-type classification. Participants were removed from further analysis if they scored less than 50% on the *no event* control comparison trials. Two participants were removed from analysis via this method.

5.3.1 Solution rates

Figure 5.1 shows mean solution rates for Complexity and Event-type by Set and Cost. A repeated measures ANOVA was conducted on the proportion of correct responses across Complexity, Event-type, Mental set, and Cost. No significant difference was found between set ($M=0.87$, $SD=0.13$) and no set conditions ($M=0.86$, $SD=0.11$), $F(1,56)=.056$, $p=0.82$, $\eta^2=.001$, 95% CIs [.83, .92] and [.81, .92]. A main effect was found for Cost, with those in the cost group ($M=0.92$, $SD=0.08$) solving significantly more trials correctly than those in the no

cost group ($M=0.82$, $SD=0.15$), $F(1, 56)=7.14$, $p=0.01$, $\eta^2=.11$, 95% CIs [.87,.96] and [.77,.87]. A main effect was found for Complexity, with complex trials ($M=0.78$, $SD=0.21$) solved significantly less often than simple trials ($M=0.96$, $SD=0.11$), $F(1,56)=55.20$, $p<0.001$, $\eta^2=.50$, 95% CIs [.72,.83] and [.93,.99]. Another main effect was found for Event-type, with conflict trials ($M=0.93$, $SD=0.11$) being solved at a higher rate than environmental trials ($M=0.81$, $SD=0.20$), $F(1,56)=29.10$, $p<0.001$, $\eta^2=.34$, 95% CIs [.90,.95] and [.76,.86].

A significant interaction was found for Complexity by Event-type $F(1,56)=56.00$, $p<0.001$, $\eta^2=.26$. Complex environmental trials ($M=0.68$, $SD=0.32$) had a lower solution rate than complex conflict trials ($M=0.88$, $SD=0.15$), simple conflict trials ($M=0.98$, $SD=0.09$) and simple environmental trials ($M=0.94$, $SD=0.14$). Another significant interaction was found for Complexity by Cost $F(1,56)=6.81$, $p<0.05$, $\eta^2=.11$. Complex trials were solved less often by the no cost group ($M=0.69$, $SD=0.23$) than by the cost group ($M=0.86$, $SD=0.16$), with no difference between groups for simple trials ($M_{\text{cost}}=0.98$, $SD=0.07$), $M_{\text{no cost}}=.95$, $SD=0.13$). A three-way interaction was found between Set, Complexity and Cost, $F(1,56)=4.65$, $p<0.05$, $\eta^2=.08$. Inspection of Figure 5.1 below shows significantly lower solution rates for complex trials for the no-set group in the no-cost condition than all other conditions.

All remaining interactions were not significant. This includes the two-way interactions of Complexity by Set, $F(1,56)=0.16$, $p=0.69$, Event by Set, $F(1,56)=0.99$, $p=0.32$, and Cost by Event, $F(1,56)=1.73$, $p=0.19$. The three-way interactions of Event by Set by Cost, $F(1,56)=1.58$, $p=0.21$, Complexity by Event by Set, $F(1,56)=0.14$, $p=0.71$, and Complexity by Event by Cost, $F(1,56)=3.65$, $p=0.061$. The four-way interaction for Complexity by Event by Cost by Set was also not significant, $F(1,56)=2.885$, $p=0.09$.

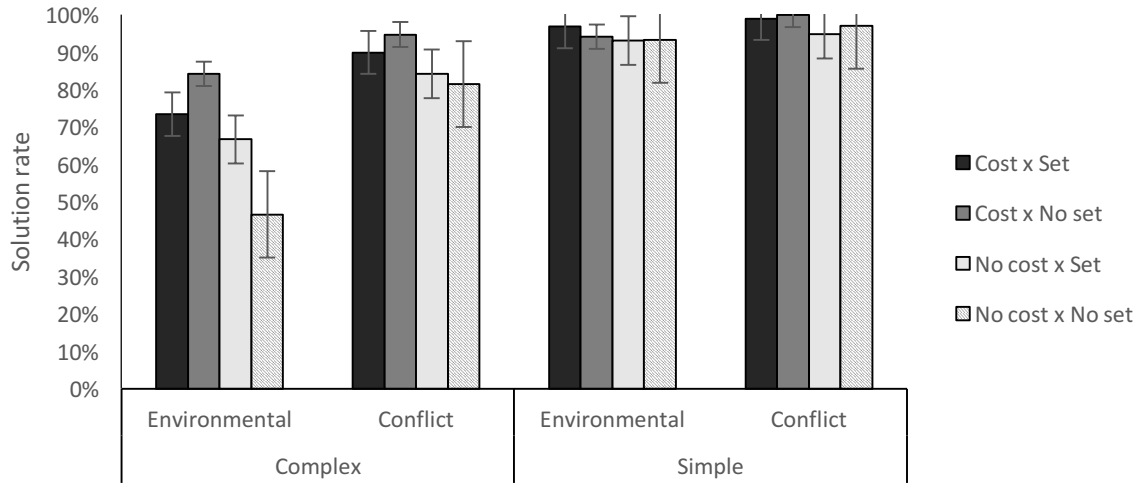


Figure 5.1. Mean proportion of correct solutions for complexity and event-type, by set and cost ($n=68$).

5.3.2 Solution times

Mean solution times for Complexity and Event-type by Cost and Set are shown in Figure 5.2.

A repeated measures ANOVA found a significant main effect for Complexity, with complex trials ($M=24.7s$, $SD=9.3s$) taking longer than simple trials ($M=22.6s$, $SD=7.9s$), $F(1, 60)=6.87$, $p<0.05$, $\eta^2=.10$, 95% CIs [22.4,27.0] and [20.5,24.6]. A significant main effect was found for Event-type, with conflict trials ($M=20.6s$, $SD=6.13s$) faster than environmental trials ($M=26.6s$, $SD=11.0s$), $F(1,60)=38.33$, $p<0.001$, $\eta^2=.39$, 95% CIs [19.0,22.2] and [24.0,29.4]. The main effect for Cost was also significant, with the no cost condition ($M=21.5s$, $SD=7.31s$) faster than the cost condition ($M=25.8s$, $SD=8.14s$), $F(1,60)=4.67$, $p<0.05$, $\eta^2=0.7$, 95% CIs [18.5,24.4] and [23.1,28.5]. The main effect for Set was not significant, $F(1,60)=0.16$, $p=0.69$.

A significant interaction was found between Event-type and Cost, $F(1,60)=5.87$, $p<0.05$, $\eta^2=.09$. Inspection of Figure 5.2 indicates that the cost group took longer to solve environmental problems ($M=30.0s$, $SD=11.26s$) than the no cost group ($M=23.3s$, $SD=9.69s$). A significant interaction was also found for Complexity by Event, $F(1,60)=18.39$, $p<0.001$,

$\eta^2=.24$. Simple conflict trials were solved more quickly ($M=17.76s$, $SD= 5.77s$) than simple environmental trials ($M=27.36s$, $SD=11.45s$), with little difference between complex environmental trials ($M=25.94s$, $SD=12.24s$) and complex conflict trials ($M=23.50s$, $SD=9.64s$).

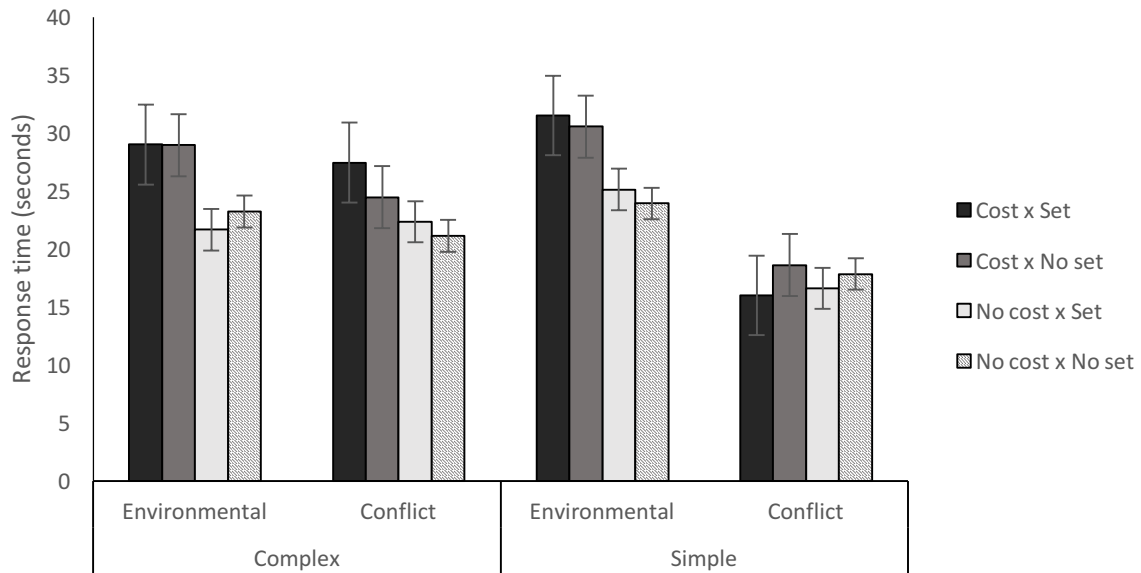


Figure 5.2. Mean solution times (in seconds) for complexity and event-type by cost and set ($n=68$).

A three-way significant interaction was found between Complexity, Event-type and Set, $F(1,60)=7.06$, $p=0.01$, $\eta^2=.11$. Examination of Figure 5.2 suggests that those in the set condition were significantly faster to respond to simple conflict trials than all other conditions. Another three-way interaction was significant between Complexity, Event and Cost, $F(1,60)=4.50$, $p<0.05$, $\eta^2=.07$. Solution times were slower for the cost group in when solving complex environmental problems than all other conditions.

Two-way interactions that were not found significant included Complexity by Set, $F(1,60)=1.27$, $p<0.26$, Complexity by Cost, $F(1,60)=2.88$, $p<0.09$, and Event by Set, $F(1,60)=0.01$, $p<0.97$. Three way interactions that were not significant were Complexity by Set by Cost, $F(1,60)=0.07$, $p<0.79$, Event by Set by Cost, $F(1,60)=0.02$, $p<0.90$, and

Complexity by Event by Cost, $F(1,60)=2.71, p<0.11$.

The four-way interaction for Complexity by Event by Set by Cost was also not significant, $F(1,60)=0.08, p=0.78$.

5.4 Discussion

Results from the previous experiments suggest that introducing a mental set inhibits solution rates on a complex trial when the set trial type and complex trial type share surface similarities (Experiment 1). Introducing feedback eliminates the detrimental effect of mental set (Experiment 2), but receiving feedback does not necessarily lead to participants using optimal strategies, but instead using heuristic recognition strategies that were better refined with feedback. Experiment 3 examined whether introducing cost of failure would promote a change to an optimal (inference-based) strategy.

The results show that imposing a cost for failed trials does lead to higher solution rates. This improvement was also accompanied by an increase in solution times for those in the cost group compared to the no cost group, suggesting that when there is a potential for a penalty it encourages people to alter their problem-solving strategy to a slower but more accurate inferential strategy. However, it is still not clear what the nature of this change is. It could be that cost makes participants take more care, perhaps double-checking their reasoning so that they are more likely to arrive at the correct solution or pick up on errors. Alternatively, participants may employ more analytical strategies (such as logical reasoning, or algorithm generation) that require more cognitive resources, thus taking more time to complete the tasks, as there is now an incentive for increased attention and effort.

The interaction between Complexity and Cost may provide support for the latter explanation. Participants in the cost conditions showed higher solution rates for complex trials than those in the no-cost conditions, suggesting they used a more optimal strategy and

switched away from a default recognition strategy on these trials but not on simple trials. This indicates that by default, recognition heuristics were used to solve most problems, and even in the presence of a cost constraint, this strategy is sufficient for success. With complex problems, the same strategy is used unless the cost of failure is high, in which case participants were able to switch to an inferential strategy.

An unexpected finding was the three-way interaction between set, complexity and cost for solution rates: Fewer complex problems were solved in the no-cost condition for the no set group. It could be that establishing a mental set confers a small benefit to problem solvers using recognition-based heuristics (in the no-cost condition where such strategies are suggested to dominate) for the detection of a change – this may be due to those experiencing the set become familiar with the problem-type used to establish the set. The idea of mental set is analogous to a ‘muscle memory’ for cognitive operators. By exposing participants to the same stimulus, they may become adept at recognising it, and then by extension, better at recognising when a new stimulus is not a match – a form of change-detection heuristic.

Further evidence for the notion that mental set may confer a bonus to problem solvers exists in the solution time data. The set group responded significantly faster than those in the no-set group for simple conflict trials. This suggests that those in the set group are better at recognising simple conflict trials. This may benefit solution rates on complex problems by introducing a more rapid conflict/not-conflict classification. During the onset of a complex environmental trial, those in the set group are perhaps more aware of what a conflict looks like and notice subtle differences in input conditions, whereas those in the no set group do not pick up on the smaller differences and instead view the output conditions as the most diagnostic information. This could help explain why for the no cost group, those in the set condition were better at solving complex problems and explain the significant interaction.

The key findings from this experiment are that introducing a cost for incorrect

solution improves solution rates. This could be due to selecting a more optimal strategy although it is not possible to determine what type of strategy this may be. The current experiment had an unequal number of each trial type. This is done for a number of reasons. Firstly, the primary research question is about anomaly handling and one of the defining characteristics of an anomaly is its low frequency. If anomalies were frequent they would just be another problem type. One specific anomaly (complex environmental event) occurs only once in each block of six problems. The mental set that precedes this anomaly consists of four of the same simple conflict trials. Therefore, there will always be four times the number of simple conflict events as complex environmental events. To balance this lack of parity, six simple environmental event trials were used. This number is the same as the number of complex environmental events and can be used to check the effect of frequency. If low solution rates occurred for both simple and complex environmental events then it would suggest that their low frequency could be responsible. However, in our experiments simple environmental (6 trials) and simple conflict (24 trials) have similarly high solution rates while complex environmental trials (6 trials) have significantly lower solution rates, suggesting that poor performance is due to complexity rather than frequency. This again reaffirms the previous research on cues as a valid measure of creating a paradigm in which complex problems exist using minimal information by designing the problems in such a way that cue-dependency can effect solution rates across different types of problem. However, it would be interesting to find out what happens when the cues weighing in favour of the problem-type used to establish the set is different to the cues used for the majority of trials, which is the basis for the next experiment.

6 Experiment 4: Manipulating mental set

6.1 Introduction.

Experiments 1, 2 and 3 had unequal numbers of each trial type. As outlined in the discussion of Experiment 3, this was necessary for the creation of anomalies, which by their very nature must be lower in frequency than standard events. One potential issue arises from this around the establishment of mental set, which is achieved using a series of similar event types. In the experiments reported so far, the event type used to establish the mental set was the same event type that was also the most frequently occurring.

However, given that mental set may induce a familiarity with simple conflict events that may actually improve detection of complex environmental trials (seen in Experiment 3), there has not yet been a thorough attempt to separate the effects of mental set from the impact of stimuli familiarity. Experiment 4 was conducted to alter two factors that have remained constant in previous experiments. The first was to reduce the overall number of simple conflict trials, in order to reduce the potency of a single recognition heuristic strategy while still testing the effects of mental set. The second was to reduce the familiarity of simple conflict trials, by using complex conflict problems to create the mental set. Complex conflict trials introduce more noise and so are less easy to form a stable prototype of what a conflict looks like in the set condition. This may reduce the ability to solve complex problems in participants who attempt to solve complex problems by noticing the onset of change rather than utilising more analytical strategies. So far, a single cue, the weapons data, has been sufficient to successfully complete the majority of problems. By changing the type of problem used to establish set we can eliminate a single cue from being as important.

Besides the introduction of an additional level of set, a further difference in this study was an adjusted base-rate for trial types. In previous experiments, the trial type used to create mental sets also occurred with the highest frequency. In this study, the base rates of trial types

were changed so non-set trials were at a higher frequency than trials used to establish sets. This change was introduced to determine whether the impact of a repetitive consecutive series of trials of the same event type is truly one of mental set, and not just one of stimuli familiarity. Previous experiments used conflict event trials to establish a set in a paradigm where conflict trials also represented the largest number of trials to be solved. In Experiment 4, conflict events were used to establish the mental set but the number of environmental trials was also increased. This change allowed the differences in strategy used to solve set- trials or to solve the most frequently occurring trial type to be explored.

The structure of this experiment allows a number of hypotheses. Firstly, when the trial-type used to establish set requires more cues to be used, then this will make it harder to distinguish the onset of a complex trial. In this case, when complex event trials are used to establish a set, we expect solution rates on the subsequent complex environmental event trials to be lower than compared to a set of simple conflict event trials.

Secondly, as discussed in Experiment 3, being exposed to similar stimuli repeatedly may not only establish a set, but also allow that set to confer the ability to better identify problems of the same nature. Therefore, when there is no set established, it was expected that participants would take longer to solve simple trials, and also experience reduced solution rates as a result of not having the sustained familiarity with a particular trial type that can be used to identify a correct solution.

6.3 Method

6.3.1 Participants

Participants were recruited online through social media and online advertisement. A total of 46 participants were recruited for this study (17 female, 29 male). Age of participants ranged from 18 to 54 ($M = 25.77$, $SD = 8.7$). Online recruitment extended the geographical

reach of this research; participants came from the United Kingdom (n=18), United States of America (n=14), Australia (n=3), Canada (n=3), Germany (n=2), Jamaica (n=1), Mexico (n=1), New Zealand (n=1), Poland (n=1), and Russia (n=1). All participants were fluent in English. Participants were explicitly instructed to not take part in this research if they had familiarity with the paradigm by completing previous online research (Experiment 2). A further 61 participants began the experiment but did not finish, these participants were considered to have withdrawn and were excluded from analysis.

6.3.2 Design

A 2x2x3 mixed design was used for this study. A between-subjects factor of *Set* had three levels (*no set*, *simple set*, *complex set*). Within-subjects factors of *Complexity* (*simple trials*, *complex trials*), and *Event-type* (*conflict event*, *environmental event*) were as in previous experiments.

6.3.3 Materials

The system, rules and participant interactions were the same as in previous experiments, and only the composition of trials was changed. In this study, 54 trials were presented across 3 problem blocks. A problem block consisted of 18 trials that culminated with a *complex environmental event* trial. In the *simple set* group, this end trial was preceded by five *simple conflict event* trials (similar to how mental set was established in previous studies). In the *complex set* group, the end trial was preceded by five *complex conflict event* trials. In the *no set* group, the final trial in each problem block was preceded by a variety of all possible trial types.

Trials of the *simple set* and the *no set* groups were identical but in different orders. These were 15 *simple conflict event* trials, 24 *simple environmental event* trials, 6 *complex conflict event* trials, 6 *no-event* trials and 3 *complex environmental event* trials. The *complex set* group had *complex conflict event trials* to establish set rather than simple conflict event

trials, for a final composition of 15 *complex conflict event* trials, 24 *simple environmental event* trials, 6 *simple conflict event* trials, 6 *no event* trials, and 3 *complex environmental event* trials. The specific order of these trials can be found in Appendix 2.

6.3.4 Procedure

The experimental procedure was the same as in Experiment 2. The study was hosted online using a researcher-owned web address. This web address was circulated online among academic interest groups in order to attract suitable participants. Participants visiting the web address were welcomed, presented with an information sheet and consent form. Participants were informed that participation was voluntary and that they could exit the study at any time without repercussion by closing the web browser. Participants were then presented a set of instructions that inform the participant about the system they are about to use. Rules were explained, examples were given, and participants were allowed to explore this until they felt ready to begin (Same materials as found in Appendix 1). Once participants were ready they could begin the experiment. After each decision was submitted, the participant received feedback informing them of the correct response to each problem.

6.4 Results

Data include the categorisation decision made by the participant for each trial, and the time taken to make that decision. A trial was considered correct if the participants' categorisation was consistent with the trial event-type classification. For the purpose of analysis, answers were converted into solution rates for each trial type.

6.4.1 Solution rates

Figure 6.1 shows the proportion of correct responses for each condition in Experiment 4. A mixed ANOVA was conducted on the proportion of correct responses across Event-type, Complexity, and Set. No significant main effect was found for Set, $F(1,43)=1.67$,

$p=0.2$. A significant main effect was found of Complexity, with simple trials having a higher solution rate ($M=0.96$, $SD=0.12$) than complex trials ($M=0.83$, $SD=0.22$), $F(1,43)=32.91$, $p<0.001$, $\eta^2=.43$, 95% CIs [.93,.99] and [.76, .89]. A significant main effect was also found for Event-type, with conflict events being solved more often ($M=0.95$, $SD=0.14$) than environmental events ($M=0.84$, $SD=0.21$), $F(1,43)=24.07$, $p<0.001$, $\eta^2=.36$, 95% CIs [.91,.99] and [.78, .90].

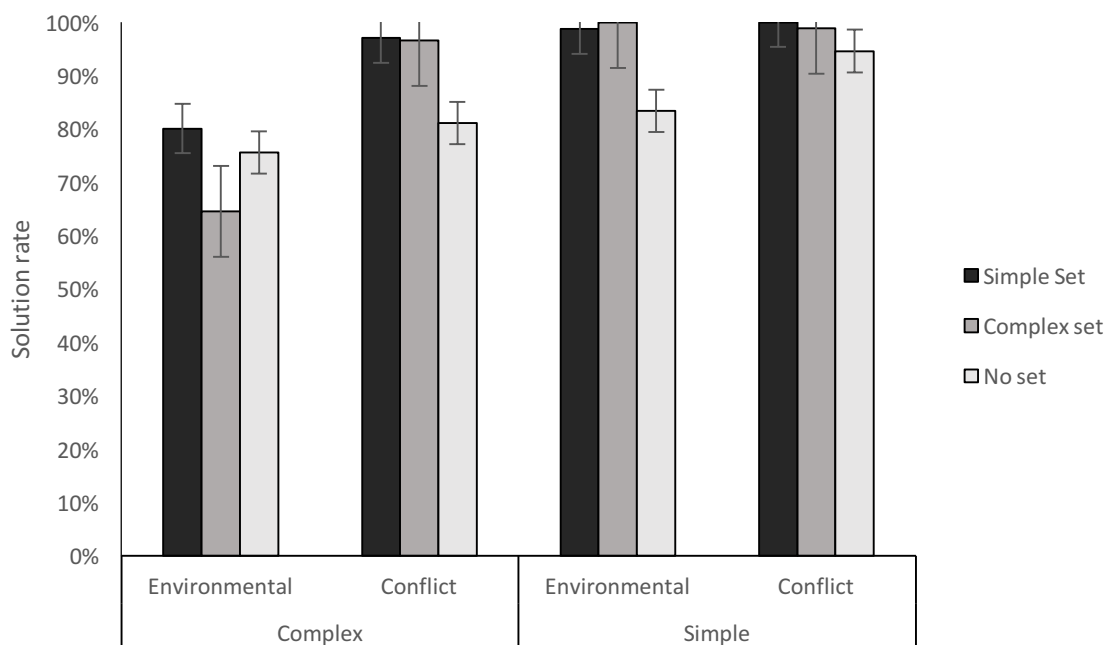


Figure 6.1. Mean proportions of correct solutions for Complexity and Event type by Set ($n=46$).

A significant interaction was found between Complexity and Event-type, $F(1,43)=5.91$, $p<0.05$, $\eta^2=.12$. Examination of Figure 6.1 suggests complex environmental events ($M=0.73$, $SD=0.31$) had lower solution rates than complex conflict events ($M=0.92$, $SD=0.23$), while solution rates did not differ for simple environmental events ($M=0.94$, $SD=0.20$) and simple conflict events ($M=0.98$, $SD=0.06$). A three-way interaction was also

found between Complexity, Event type and Set , $F(2,43)=3.75, p<0.05, \eta^2=.15$. Figure 6.1 indicates this interaction is due to reduced solution rates for both set conditions for complex environmental events.

All other interactions were not significant. These were the two way interactions for Complexity by Set, $F(2,43)=1.30, p<0.29$, and Event by Set, $F(2,43)=1.16, p<0.32$.

6.4.2 Solution times

A mixed ANOVA was conducted on the solution time data (means shown in Figure 6.2). A significant main effect was found for Complexity, with complex trials ($M=24.1s, SD=8.27s$) being slower than simple trials ($M=20.5s, SD=9.08s$), $F(1,44)=10.65, p<0.005, \eta^2=.20$, 95% CIs [21.62, 26.57] and [17.78, 23.17]. Main effects for Set ($F(1,44)=0.40, p=0.6$), and Event-type ($F(1,44)=2.95, p=0.09$) were not significant. There were no significant interactions. These interactions were Complexity by Set, $F(1,44)=0.35, p<0.71$, Event by Set, $F(1,44)=1.89, p<0.16$, Complexity by Event, $F(1,44)=0.01, p<0.98$, and Complexity by Event by Set, $F(1,44)=2.45, p<0.10$.

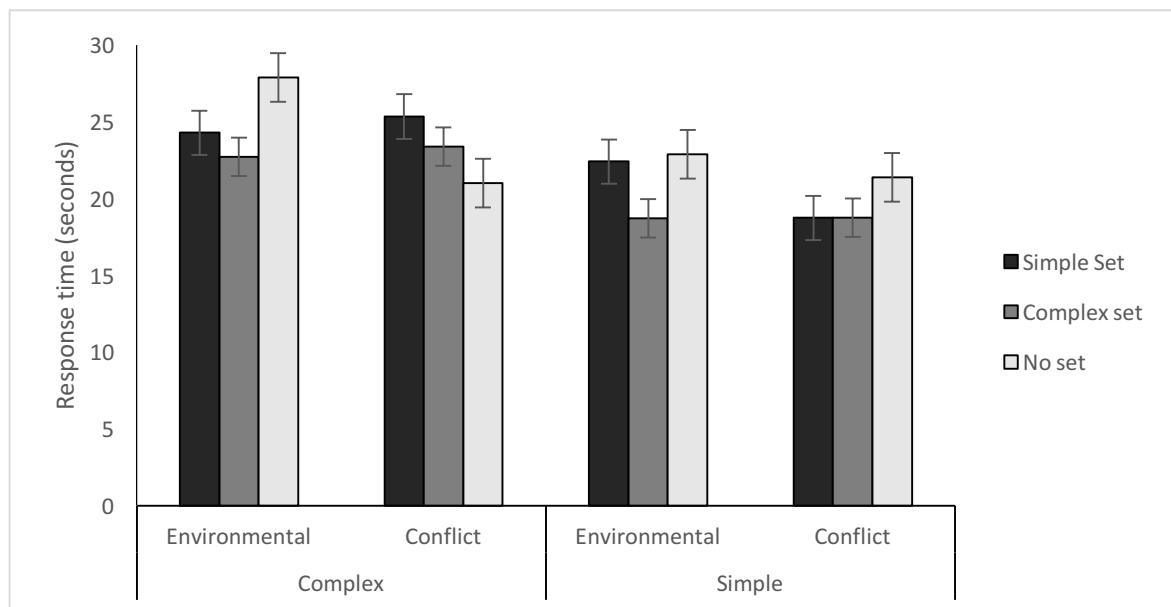


Figure 6.2. Mean solution time (in seconds) for complexity, event type and set type ($n=46$).

6.5 Discussion

Previous experiments established that complex trials are more difficult and take longer to solve than simple problems (Experiments 1, 2 and 3). Experiment 1 also indicated that exposure to a mental set (established with a series of concurrent and consistent trial types) reduced solution rates for complex trials when the trial type used to establish the set, and the complex trial type share surface similarities. In Experiments 1-3, set was established using simple conflict problems. Similarity with simple conflict trials is seen in the complex environmental trial, as both problem types have the same identical final distribution of weapons.

Results from Experiment 4 are consistent with previous findings. Solution rates were lower and solution times were longer for complex than simple trials. Participants were less successful at correctly identifying environmental than conflict events, and complex environmental events were hardest to identify. In this experiment, the complex set condition exhibited lower solution rates than the simple set condition for the complex environmental event trial, supporting our hypothesis that the more similar the set-making trial type is to the subsequent different problem, then the more likely that an error would occur. In this study, the simple set condition was consistent with previous implementations of mental set, where a series of consecutive simple conflict events were used to establish mental set prior to the trial type that shares similarity, the complex environmental problem. In the complex set condition, complex conflict event trials were used to create the mental set prior to presentation of a complex environmental event. The impact of this manipulation was significantly reduced solution rates for complex environmental events compared with complex conflict events in the complex set condition.

This result can be explained using the same surface similarity argument given in the

previous chapter for the initial creation of mental sets. Previous experiments indicated that establishing a mental set affects solution rates due to similarities between the set trial type and the subsequent different trial that possesses similar characteristics. By presenting a series of trials of the same type, the participant may use recognition heuristics to classify the trials they are familiar with. This is efficient because the simple trial types have distinct surface details that can be used to separate and classify (two very different weapon distribution profiles that are immediately apparent between conflict events and environmental events). However, the complex trial types are functionally different from the set trial type, although it does share some surface features with the set trial type. Solution rates in set conditions in previous experiments were reduced, we propose, because participants used surface features as categorisation criteria and thus encounter problems only for the target trial type.

Surface similarity on our trials can be described as a binary status of ‘hit’ or ‘miss’ that comes from the two main cues within the task; input data (environmental data) and output data (weapons data). A ‘hit’ would be defined as a relevant change within either data set that would have consequence on the classification of the trial. For example, simple conflict event trials have a ‘hit’ in the output data (a sharp increase in weapons), and a ‘miss’ in the input data (no environmental metrics make a discernible change). Simple environmental event trials act oppositely and have a ‘hit’ in the input data and a ‘miss’ in the output data. When using simple conflict event trials to establish a set, it could be that participants begin to look for these hits in the output data only, and limit their strategy to using a single cue, which would then result in problems when facing complex environmental event trials which possess a ‘hit’ in both input and output data. If a participant with an established set was using the ‘hit’s in the output data only, they would fail to differentiate these trial types leading to the observed lower solution rates for complex environmental event trials. Complex conflict trials also possess a ‘hit’ in both input and output data, as they

contain both a drastic increase in weapons indicative of a conflict, but also have movement within the environmental metrics. This additional similarity may explain why establishing a set using complex conflict event trials is more detrimental to performance on complex environmental event trials, as they experience even more surface similarity than simple sets.

A limitation of previous studies has been that the trial type used to establish mental sets has also been the predominant trial type in terms of frequency of occurrence. This leaves a question as to whether mental set or familiarity, derived from being the most common trial type, has the most effect on the reduced solution rates to complex problems. This study addressed this concern by changing the base rate for frequency of occurrence for trial types. Previous studies established sets with conflict event trials, and as a result, conflict event trials were the majority of trials. In this current study, conflict event trials were retained as the establishing set trial type, but problem blocks were expanded, and environmental event trials were made the most occurring trial type.

Previous experiments have found that not being exposed to a mental set gave higher solution rates, a result attributed to the lack of interference from mental set. However, this study found that the no set condition solution rates were lower for simple trials compared to both set conditions, lower for complex conflict event trials compared to both set conditions and were lower than simple set for complex environmental event trials. This can be interpreted as both a barrier to success for the no set condition, and as the set conditions conferring a benefit and facilitating problem solving. This notion of mental sets facilitating problem solving was first encountered in Experiment 3, where advantages in solution times were found for the set condition in identifying simple conflict trials. It was argued that this benefit is mainly conferred to identifying set-consistent trials. In the current study, both set type conditions have improved solution rates over the no set condition for simple environmental event trials, simple conflict event trials, and complex conflict trials. This

benefit could be due to the familiarity with conflict event trials that are exposed to participants in succession during the set establishment and the subsequent development of prototypes. The set group are also familiar with the environmental event trials due to their high rate of occurrence. This allows those in the set conditions to establish a quick recognition-based dichotomy between conflict and environmental events that allows more accurate classification of simple trials. Those in the no set condition do not experience or gain familiarity with conflict event type problems and so do not have a repertoire of these trial types. Conversely, complex set, while appearing to confer a benefit in the instance of simple trials and set-congruent trials, also suffers in the complex environmental event trials as reasoned above due to surface similarity.

Overall, this study suggests that familiarity with event-types may be an important factor on successful solution of trials, and that experiencing a mental set may facilitate this familiarity. This can be seen in improved solution rates for some trial types in both set conditions. However, the complexity of the set has an effect on complex trial solutions if the set trial types have surface similarity with the target complex trial. It was found that when the set trial type, and the target complex trial type that followed the set, shared more similarity, then solution rates decreased. This is manifested by the complex set condition have the lowest solution rates for the target complex environmental event trial.

7 Experiment 5 –Protocol analysis of strategies

7.1 Introduction

Previous experiments have shown the influence of factors affecting solution rates to complex problems presented using the anomaly handling paradigm. Changes in behaviour were elicited by the manipulation of feedback, mental set, cost, and problem frequency. These effects have been discussed in terms of the strategies used by participants, and while strategic differences have been supported by the data, they themselves have not been subjected to objective scrutiny. For example, Experiment 2 introduced feedback and found that this greatly improved solution rates. However, it remains unclear whether it was due to a switch away from heuristic strategies found to be inadequate, or through an improved recognition ability. While these were discussed there was no single approach supported by the evidence. Similarly, Experiment 3 introduced cost of failure, and showed that while this promoted a switch to a more efficient strategy, what these strategies were was unclear.

In this chapter, a descriptive account of what strategies participants were using, and how these correlate with success, was developed through analysis of participant protocols. This study was an investigation into the strategies are being used and whether these changed based on the types of problem encountered.

One such method for capturing the descriptive nature of what each participant is doing when they are completing the task is the ‘think-aloud’ method of capturing verbal protocols (Ericson & Simon, 1998). This is where participants are prompted to think out loud and verbalise their thoughts while simultaneously completing the task, in order to obtain information around what aspects of the simulation is being attended to and what information elements contained within the simulation are being used during decision-making. This methodology has also previously been used successfully to explore reasoning (Ball, Ormerod, & Morley, 2004). This provides valuable insight into how participants go about completing

the task that has not previously been studied in such detail. By making participants verbalise their thoughts as they complete the task, the aim was to examine the separate cognitive operations participants use when navigating the data elements contained within the simulation. By examining data at such a level, we can create a series of hypotheses around how such cognitive operations are organised.

Firstly, it was predicted that the number of cognitive operations exhibited by participants in undertaking complex problems would be greater than simple problems. Complex problems by their nature require more thinking to arrive at the correct answer, so confirming this hypothesis indicates the validity of using verbal protocol methodology and evidence that participants are able to accurately report on their thoughts during completion of the task.

Secondly, it was predicted that solutions to conflict problems would show fewer cognitive operations than environmental problems because the data structure of an environmental problem is different to that of a conflict problem. As explored in the discussion of Experiment 2, our simulation can be described as two data streams – ‘input’ (environmental data) and ‘output ‘weapons data’. A conflict problem can be represented by a disturbance within one data stream (output), whereas environmental problems have disturbances in both data streams. While the logical complexity of simple conflicts and simple environmental problems are the same, it may be that the differences in data structures necessitate additional cognitive operations for the identification of environmental problems.

Thirdly, there will be consistencies in the order that cognitive operations are used across multiple problems that will allow the mapping of a general problem-solving process. It is expected based on results of previous experiments that participants may be using heuristic strategies. These strategies will be possible to identify with the use of verbal protocol analysis. If these strategies are taking place, then it would be expected that some participants

will refer to single cues during problem completion and will not refer to other cues that are required for inferential reasoning.

7.2 Method

7.2.1 Participants

Four volunteers were recruited from the local area. Ages ranged from 25-33 ($M=27.8$, $SD=3.6$), with three males and one female.

7.2.2 Design

A 2x2 within-subjects design was used for this study. Within-subjects factors of *Complexity* (*complex trials*, *simple trials*) and *Event-type* (*conflict event*, *environmental event*) were retained. As with all previous experiments, this created four distinct trial-types; *simple conflict event*, *simple environmental event*, *complex conflict event*, and *complex environmental event*.

7.2.3 Materials

The task used in this study was the same computerised version of our anomaly-handling paradigm used in previous experiments. The paradigm is a classification task where participants must evaluate each trial and categorise it based upon event-type (conflict or environmental).

Instructions were included within the system that detailed how to complete the task and explained the rules of the system. There were a total of 54 trials. These 54 trials consisted of 24 *simple conflict* trials, 6 *simple environmental* trials, 12 *complex conflict* trials, 6 *complex environmental* trials and 6 *no event* trials. A Samsung S6 mobile recording device was placed on the table to ensure capture of the participant's verbal protocol.

7.2.4 Procedure

Each participant was seated at a desk containing a desktop computer and the recording

device. Information sheets were distributed to participants explaining the nature of the experiment before gaining their informed consent. Once consent had been obtained, participants were directed towards the computer, where a computerised tutorial explained the nature of the task, the rules of the simulation, and how to record responses. Once participants had completed this section, the participants were given advice on verbalising their thoughts. It was explained that the experimenter would give prompts if the participant fell silent for more than 5 seconds. After this advice, participants were given an opportunity to practice the verbalisation process by making a drawing on the computer whilst verbalising. This gave an opportunity for participants to practice verbalisation and gain an understanding of what was expected. After this practice, each participants was given the opportunity to ask questions to the researcher. After this, the voice recorder was switched on and the participant was directed to begin the first task. During the task, participants were given feedback of the correct solution after submitted each trial.

7.2.5 Verbal protocols and cognitive operations

After the study was completed, the voice recordings were transcribed by the researcher. This process involved listening to the recordings, and creating a written version of what each participant said during their verbalisations while completing of each of the 54 individual trials. The researcher manually defined the boundaries for each trial, and then like-for-like transcribed the participant's response for that trial.

Once all transcriptions were completed, the researcher read and familiarised themselves with the resulting transcripts and generated a list of cognitive operations that were performed by participants. The list of cognitive operations resulting from the scanning and familiarisation process were:

Cognitive Operation	Description	Example
Scan data stream	Scanning or reference to examination of data in the environmental and weapons streams	“Conditions look stable” “Looking at the weapons now” “By looking at the weather”
Identify change	Identification or reference to finding a significant change in data streams	“I see a spike in Zone A” “The river is flooded”
Reference magnitude	Reference to understanding the scale of a change found	“Weapons in B have quartered” “Zone C and D have halved”
Create hypothetical	Reference to the relationship between rules and how it should influence the system / data streams	“If the mountain pass freezes, then A should go down...”
Test hypothetical	Search for information to test an observed hypothetical	“...A has gone down as expected...”
Seek verifying information	Search for additional information that reinforces an already held belief	“...and we would then see B and C go up, which they do”
Discover a violation of expectation	Find information that runs contrary to what was expected	“So B and C should go up, ...but C hasn’t...”
Generate solution	Express an understanding of the nature of the trial	“...So that must mean there is a conflict in Zone B”

Once the cognitive operations were identified, the verbalisation of each trial was coded for the operations present and the order in which they occurred. This allows us to see how many operations were conducted for each individual trial. During the coding of the cognitive operations, both the “Scan data stream” and “Identify change” were only coded once for each data stream (environmental conditions, weapons).

7.3 Results

7.3.1 Cognitive Operations

Each trial’s verbalisation was coded for the number of operations present. This allows the examination of the number of cognitive operations present during the solution of each individual trial. Of the 216 trials recorded, only 167 resulted in a verbal protocol. On some trials participants did not provide a verbal protocol because they immediately solved the problem without verbalisation, or recordings were indecipherable due to equipment failure.

A t-test was conducted on the number of cognitive operations recorded for simple versus complex problems. A significant difference was found in the number of operations present for simple problems ($M=4.95$, $SD=1.26$) than complex problems ($M=6.76$, $SD=1.44$), $t(166)=8.32$, $p<0.001$. The number of cognitive operations was also significantly lower for conflict events ($M=5.33$, $SD=1.53$) than for environmental events ($M=6.36$, $SD=1.48$), $t(166)=3.16$, $p=0.001$. Table 6 shows the mean number of cognitive operations used for each trial type.

Table 6. *Average number of cognitive operations by complexity and event type (SD)*

	Simple	Complex	Average
Conflict	4.66 (0.98)	6.81 (1.51)	5.33 (1.53)
Environmental	6.10 (1.58)	6.67 (1.33)	6.36 (1.48)
Average	4.95 (1.26)	6.76 (1.44)	

7.3.2 *Analysis of failed decisions*

Overall solution rates were extremely high (95%) and due to the verbalisation process, solution times were not included. However, it is possible to examine qualitatively the incidents of incorrect problem solving. A total of nine trials across all participants were responded to incorrectly. These consisted of three *complex environmental* trials, five *complex conflict* trials, and one *simple conflict* trial.

The operations used on the failed simple trial were “Scan data (weapons) -> Scan data (environmental) -> Generate solution”. This shows that the participant failed to notice the abrupt change in the data streams, rather than misattributing the data change to an incorrect cause. Errors such as these can be deemed to be perceptual, a prominent factor discussed in Experiment 1. Not all significant changes in the data streams are of a similar magnitude: for example, Rainfall increases by over 500%, Temperature changes from a positive to a negative scale, and so these changes are easier to spot. However, seismic activity and changes in weapons merely double in size, and while both of these values double, there is still difference in absolute terms, for example, weapons will double from 400 to 800, whereas seismic activity data will double from 3 to 6. While the scale of the change is the same, it should be easier to identify the change in weapons due to the increase in absolute values. While it is difficult to study this at a cognitive level, it is worth bearing in mind that not all errors are due to utilising incorrect problem-solving strategies, and that sometimes participants weren't aware there was a problem to solve.

Moving on to the failure to correctly identify complex conflict trials, all five began with the participants scanning the weapons data and identifying the increase in weapons in one zone. This was followed up by the subsequent scanning of the environmental data and also identifying the increases in particular environmental conditions. Where participants then

failed is in the incorrect attribution that the identified changes in environmental conditions were responsible for the resulting increase in weapons. This is incorrect because there is no possible logical relationship between those particular input conditions that could create those output conditions. By examining specific protocols we can see this misattribution being made. For example;

“Spike in Zone C, just over 50% drop in Zone B. I see large seismic activity, Zone D would have dropped about the same...and it has, Zone A has come down in line with its normal thing, so this one’s going to be environmental”

In this example, the participant scans the weapons data stream and correctly identifies an increase in Zone C (*“Spike in Zone C”*). The participant then references the size of the reduction in another zone (*“...just over 50% drop in Zone B”*). The participant now scans the environmental stream and identifies a change in the seismic activity (*“I see large seismic activity”*). Here the participant then creates a hypothetical relationship between that environmental condition and its effect on weapon supply (*“Zone D would have dropped about the same”*). The participant then seeks information to test that hypothesis (*“...and it has”*). It is at this stage where the participant now makes an incorrect attribution (*“Zone A has come down in line with its normal thing”*). The participant incorrectly normalises the fact that weapons have decreased in Zone A, despite there being no environmental reason for doing so, which should indicate a conflict.

Some insight into why the participant made this error can be found in their phraseology *“Zone A has come down in line with its normal thing”*, what is meant by ‘normal thing’ in this context is not certain, what makes the participant think that decreases are expected and normal? Further insight into this can be seen in the participants answer to the trial that directly preceded the current one, where Zone A decreased due to an

environmental cause and the participant stated "...that may be environmental as Zone A is coming down [in the presence of the correct environmental condition]", it may be that the participant is now incorrectly transferring information from the previous trial into this new trial and thus allowing the misattribution.

An examination of another failed complex conflict task shows how similar misattributions can arise:

"Increases in B and the rainfall increases, so that would mean you can't get to C and D, but the temperature stays the same so that wouldn't stop you getting to A, although A is slightly higher than C and D. I think it's probably environmental."

In this protocol, the participant correctly identifies changes in both data streams (*"Increases in B and the rainfall increases"*) and goes on to establish a prediction based on relationships (*"so that would mean you can't get to C and D, but the temperature stays the same so that wouldn't stop you getting to A"*), the next operation is where the potential misattribution occurs (*"although A is slightly higher than C and D"*). The participant compares Zone A to Zones C and D. Zones C and D have been effected by both a conflict and an environmental trigger in this trial, effectively quartering, and as Zone A is only experiencing the conflict it decreases to half its usual amount. The participant compares A to C and D as a potential reason to say it's due to the environment.

However, the participant has previously established that there is nothing wrong with the environmental conditions concerned with supply to A. In this protocol, it appears that the misattribution occurs when comparing the size of the changes to other zones suffering from an environmental trigger. Could it be that when making inferences based on the relationship of the environmental data that other incorrect inferences could also be made when perceiving data moving in a similar direction? "Zone C and D have gone down, Zone C and D are caused by environment" is correct in this case, however, it is followed by the known data of

“Zone A has gone down” and is mistakenly followed up by the same conclusion of “Zone A is caused by the environment”.

Further examination of misattribution can occur by examining the incorrect answers for the complex environmental trials. Take the following protocol as an example:

“Zone B, there’s a big spike in B, A and C have dropped again, so that’s going to be a conflict in Zone B”

Here we can see that the participant has no problems in scanning and identifying the change in the weapons data stream. However, they fail to utilise any information from the environmental data stream at all. This is evidence that the participant is utilising heuristic strategies by resolving what a conflict trial ‘looks like’ and matching this recognition based model to the current data. The participant takes no time to determine the logical relationships in the environmental data to ensure the cause of the change in weapons. There is no way to determine whether there is a conflict without examination of the environmental data and the only reasonable way in which the conclusion of a conflict can be arrived at is if the participant is using information learned from previous trials in this new scenario. Here we have another example of a transfer from one trial to another, furthering the evidence of use of cognitively cheap strategies.

Not all errors were made by neglecting full information however, such as the following example:

“Increase in B, the temperature decreases and the rain increases which would mean you can’t get to A, C and D so....I think it’s....a conflict.”

Here the participant uses both weapons and environmental data, correctly identifies all relevant disturbances, creates a correct prediction based on the logical rules of the simulation, and then fails to attribute that prediction to the correct source. This is a different type of error to not fully utilising the information available, and in fact this participant predicts perfectly

what should happen to weapon supply given the environmental constraints, thus demonstrating a good understanding of the simulation. However, when it comes to delivering an answer, the participant chooses the incorrect one. This shows that there is diversity in the ways in which misattributions and mistakes can be made.

7.3.3 *Creating a strategic model to represent decision making*

Overall, there were two dominant strategies that were used by participants that were successful. The first strategy was that of scanning environmental data. Once the environmental data has been scanned, if there were no changes detected, then the weapons data would be scanned. If no changes were detected here, a ‘no change’ solution was generated. If a change was found, then a ‘conflict’ solution would be generated. However, if a change was found in the environmental conditions, then a prediction would be created about what should happen to weapon supply. If this prediction was reflected in the data, a conclusion of ‘environmental’ was given, and if the expectation was violated, and the weapons data was anomalous compared to prediction, then a conclusion of ‘conflict’ would still be generated. This strategy of ‘environmental first’ was observed only in one participant, who used it to achieve a 100% success rate across all trials.

The second most dominant successful strategy was that of ‘weapons first’. In this strategy, participants would first scan the weapons data stream and look for any change. If no change was found, then a ‘no change’ conclusion could be given. If a change is found in weapons, then the environmental data could be scanned, if no change was found here, then a ‘conflict’ conclusion could be given, but if a change in environmental conditions was found, the relationship would have to be tested, resulting in a ‘conflict’ or ‘environmental’ conclusion as appropriate.

This strategy was used by three participants and has the benefit of quickly confirming

identifying simple conflicts which make up the majority of the trials. While this strategy may be useful for solving the type of problem that is most frequent, it carries the risk that heuristic strategies may be developed that allowed the participant to take shortcuts (such as relying on the weapons information only) which would still find success on the simple trials, but encounter errors in the complex trials. These errors would include not checking all available data to create expectations of the behaviour of the system, which was found in some of the verbal protocol error analysis above.

7.4 Discussion

Overall this study adds new knowledge of how participants make decisions using this experimental paradigm that adds weight to previous theorisation around mental set. This study was the first to qualitatively examine the types of strategies being used by participants. Participants used a think-aloud method of verbalisation to create verbal protocols. These were coded for the cognitive operations conducted by participants for each trial. It was found that there were statistically significant differences for the number of operations used. Complex problems required the use of more cognitive operations than simple trials, and that environmental trials required more cognitive operations than conflict trials.

This method of verbal protocols also allowed in-depth descriptive qualitative analysis of what happens when trials were incorrectly solved. It was found that errors can arise for a number of different reasons. This includes perceptual errors, where participants do not notice a change in the data streams. While these are not as interesting as the cognitive misattributions, it is still important to know that these perceptual errors exist. Participants were sat in front of a screen in a simulated conflict scenario and were not able to correctly identify the numbers they were looking at. While this is startling, it was rare in our study, and one can only hope that experts in the field, with real risk, will not make the same mistake. There is little that can be done from a cognitive psychological standpoint to improve these

errors from being committed in the field, other than ensuring that people have adequate time to analyse data, and perhaps instead relying on distributed cognition where no one mistake can influence an entire outcome. It may be that these perceptual errors are also a result from the academic nature of research. Participants truly understand there is no real world harm from an incorrect decision and so may not use their maximum effort to ensure an correct solution. Hopefully, this would not be an issue in military, security, or other industrial cases.

A further type of error that was found was that of transfer between trials. It was shown that in one trial, the incorrect decision was arrived at from the participant's point of view by following the logical relationships within the simulation. However, it appears that the participant was using data from a previous trial in their current attempt to determine the cause of an event. This lead to inaccurate data being used in the consideration of the outcome, and thus the incorrect decision was made. This has many implications throughout this thesis. The transfer of knowledge and strategy feeds into the notion of mental set. If the specific data can be carried over from one trial to another, then the transfer of strategy seems not only possible but highly likely. This evidence of transfer also has implications for the application of the data-frame model (Klein, Moon, & Hoffman, 2006). The participant can be considered to have developed a 'frame' of reference for understanding and working with the data that had been observed until that point. The transfer of knowledge between trials would represent the repeated use of the same frame despite the introduction of new stimuli. The participant has failed to take this into account in the development of a new frame and instead applies previous understanding to the new problem, introducing error.

Another method of failure that was found within protocols in this study was that of the evidenced use of heuristics. There was an example of a trial that was answered incorrectly because not all information in the system was attended to. This can have drastic ramifications as in this instance, the parts of the system not attended to held information necessary for the

correct classification of the problem. It can be theorised that the reason the participant felt it appropriate to adopt such heuristics was that they had worked previously on simple problems where it was successful on a number of consecutive occasions, and then this strategy was transferred to complex problems (the complex nature of the problem would not be salient to the participant until receiving feedback that their response was incorrect).

This finding is also extremely crucial for the validity of this entire thesis. Up until now, the use of heuristic strategies has been hypothesised, and has been central to the idea of why mental sets are established. Where previous experiments have supported these ideas through quantitative data and suggested their presence, this study is the first time that such mental shortcuts have been captured and isolated as a cause for failure.

However, different from previous studies, this experiment had extremely high success rates. It could be that the nature of collecting verbal protocols, and having participants think out loud has influenced the way in which trials were approached. Having participants talk out loud may have made them more aware of the evidence they were using to make their conclusions against, or they may have taken more time on each problem due to the nature of having to put their thoughts into words. While these factors that may explain why solution rates were so high, the fact that errors still occurred is impressive. It may be assumed that increase in error frequency for the non-protocol studies are due to committed the same errors identified here, but on an increased scale as they never have the opportunity to ‘catch themselves’ making a mistake and correcting it.

Overall, this study adds weight to the nature of the errors that we have been discussing throughout this thesis, we have captured evidence of the use of heuristics, discovered new errors in the transference of data knowledge between trials, and confirmed that perceptual errors can and will occur in any complex system.

8 General Discussion

8.1 Aims and Overview

Overall, this thesis has aimed to accomplish two things; first, to establish a novel paradigm capable of testing anomaly handling ability without requiring domain expertise, and second, to provide insight into the factors that contribute to success in an anomaly handling task based in the world of intelligence and conflict analysis.

A brief overview of each experiment will be given prior to discussing in more detail their contribution to the objectives of this thesis, and the wider implications to the decision-making and anomaly handling communities.

For Experiment 1, a paradigm was developed where we could test the problem-solving skills of participants as they attempted to identify and interpret anomalous events in a conflict simulation. The simulation presented participants with information about the security situation of a fictional country and the participant had to classify a situation as either posing a threat or not posing a threat. Participants had to classify a number of different types of trial. Anomaly trials were more complex than the other trials. Complexity was introduced by manipulating the number of cues that had to be attended to within the information environment of the task to reach a correct solution. This follows on from previous research by Fasolo, McClelland and Lange (2005), who discovered the more cues that needed attending to, the more reported difficulty of a decision increased. This led to a hypothesis that problems that require more cues to solve correctly would be more difficult. Experiment 1 found evidence that supported this, in that complex problems (depending on more cues) had lower solution rates and slower solution times than simple problems. This shows the importance of cue dependency to the difficulty of a problem, even with a minimal number of cues available.

Experiment 1 also introduced the concept of establishing a mental set, which was

achieved via exposure to simple problems of the same type in a repetitive fashion. It was proposed that exposure to these repeated simple problems established a mental set within participants. Mental set was found to inhibit successful problem solving on subsequent complex problems when the problem had similar features (cue characteristics) to those in the problem types used to establish the set. Mental set is likely to have arisen due to the use of cognitive heuristics described in Chapter 1 (e.g., Evans, 1989; Fasolo, McLelland & Lange, 2005; Tversky & Kahneman, 1974). The most dominant form of these heuristics appears to be recognition based, where individuals compare new cues to those previously encountered. This finding supports the qualitative model of sense-making of Klein, Moon, and Hoffman (2006) in their data-frame theory, where individuals use previously experienced exemplars to aid solution of novel problems. However, whereas previous research found individuals can create adjustments from this recognition-based starting point, in Experiment 1, participants attempted to use the previously encountered solutions for solving the new problems without adjustment, and consequently failed.

Experiment 2 examined the nature of feedback on performance. When given access to direct and immediate feedback (correct/incorrect), participants were able to overcome the negative effect of mental set found in Experiment 1. Feedback appears to have affected performance due to the correction of inappropriate strategy choices. Any use of heuristic strategies would fail on complex problems and participants would be notified of this in the form of an incorrect answer. This is supported by evidence that solution rates for the first complex problem were lower than all further presented complex problems, although this change over time narrowly missed significance ($p=0.058$).

Learning over time has implications for research on mental set, as it shows that feedback is capable of breaking a set, and that once broken, mental set is not re-established for the same problem type. However, solution rates for complex problems were still lower

than those for simple problems. This means that while the mental set did not exacerbate the difficulty in solving complex anomaly problems, they were still solved less often than simple problems and so something about the nature of these complex problems still leads to lower solution rates.

Experiment 3 attempted to increase solution rates for complex problems by introducing a penalty cost for incorrect decisions. This was achieved by the reduction of financial reward to participants for each incorrect answer given. This was found to increase the number of correct answers given when compared to those participants whose reward was not dependent on performance. Moreover, effects of set not only disappeared, but in some conditions were reversed (i.e., set condition participants outperformed no-set condition participants). This result suggests that when there are real stakes in the outcome of a decision, there may be more effort put into that decision that inoculates the participant to succumbing to mental set.

The idea that there may be particular circumstances where exposure to a mental set confers a benefit when utilising sub-optimal heuristic strategies is a novel finding. One possible explanation for this unexpected result is that set encourages the rapid development of effective cue recognition when experiencing consecutive stimuli of the same type. It may be that, by detecting expected cues more effectively, minor differences (e.g., the presence of anomalous cues) are easier to recognise against a backdrop of expectation. Evidence in support of this account comes from the solution times. The usual effect of mental set is to speed up performance, since participants are applying a solution by routine. In this experiment, when not influenced by the presence of cost, solution times to the set conditions were actually slower than in the no-set conditions. This result is consistent with participants noticing an anomalous cue and switching to an inferential strategy in mental set conditions.

This account of recognition of stimuli would fit within a single ‘frame’ as described

by the RPD data-frame model, and would only require adjusting when the participant is informed their solutions were incorrect. A frame is developed which can be used in the approach to solve encountered problems. The initial frame may be simple and use the least cognitive resources, thus relying on a single salient cue, but when participants are informed their answers are incorrect, participants must alter their frame to accommodate this new information and adjust the frame to a superior one that will not fail in the same situation.

Experiment 4 aimed to identify the different contributions to successful anomaly handling of set establishment and familiarity with the stimuli type. This was achieved by changing the rate of occurrence so that the most frequent problem type and the problem type used to establish the mental set were different. Experiment 4 also examined the nature of set when established with different problem types. These changes resulted in two interesting findings. First, solution rates were higher for simple problems for those where a mental set was established compared to those without an established set. Second, the effect of mental set on complex problems was dependant on how similar the set-making and the complex problem were.

This represents further evidence that there is a benefit to set in some instances. In this case, the rapid development of cue recognition gained by establishing the set can be used to solve simple problems effectively. When there was no set established, participants were less able to correctly identify simple problems suggesting that establishment of the mental set allowed those participants a better ability to detect when subsequent cues were a match.

This effect of set establishing a strong ability for cue recognition provided a benefit on simple problems. However, this experiment also detected that when the set was established using complex problems (and thus more cues) that performance on other complex problems was lower. This can be explained by the type of cues present in the problems used to establish the set. The cues in complex problems used to establish the set (complex conflict

events) were more similar to the complex problems that suffered lower solution rates (complex environmental events). This means that the improved cue recognition ability supplied by establishing a mental set is only effective at differentiating problems based on simple recognition and is not powerful enough to aid complex problem solving. This shows that recognition heuristics can aid a switch in the strategy used to solve complex problems, but only if those recognition heuristics promote the ability to distinguish between problems. When the cues that have enhanced recognition (though mental set) are too similar to complex problems, then this strategic shift does not occur.

Experiment 5 set out to gather further evidence of the strategies used by participants, by collecting and analysing verbal protocols from participants undertaking the task. The analysis allowed counts of cognitive operations to be made. These counts revealed, as one would expect, that complex trials involved more cognitive operations than simple trials, but also the environmental trials required more cognitive operations than the conflict trials. The study had higher solution rates than the previous studies, perhaps a by-product of verbalising. Nonetheless, participants still made many errors. Three main classes were observed: perceptual, where participants failed to see or misread key pieces of data; transfer between trials, where participants used data from previous trials to solve the current trial; and inappropriate heuristics, where participants used simple heuristics for trials that required inferential decision-making.

The errors identified in Experiment 5 help explain performance on complex problems across previous experiments. Perceptual errors, where participants failed to notice the presence of cues, may be explained by the level of attention participants are applying to the task. Bettman, Johnson, Luce and Payne (1993) found that there was an association between effort and accuracy. If participants are not putting in the levels of effort necessary for the problem-solving to occur, then this will lead to failure on the task as participants are not

using required information to solve the problem. While this led to errors in Experiment 5, it also helps explain why errors were made across previous experiments. Experiment 3, where cost was introduced, gave participants a motivation to pay attention. As a result, there were increased solution rates in those participants that were financially rewarded. The presence of a financial reward itself cannot influence the strategy used to solve problems, so it may be that instead the reward encouraged participants to pay more attention and invest more effort into the task, and as a result, did not experience the same levels of perceptual errors incurred by participants that did not have a motivation for cognitive investment in the task.

Errors that demonstrated transfer between trials may be a secondary effect of mental set. When set establishes an enhanced recognition ability for a specific cue by repeat exposure, it could be that this cue is then misattributed across trials where it is not present based on the prior frequency of the cue. While this is a type of perceptual error, it is fundamentally different to that described above. The previous discussed perceptual error deals with cues that are present but are not noticed, perhaps due to lack of attention or effort, whereas this type of error deduces that participants act as if a cue is present when it is not. This may be as a result of the established mental set, where after repeated exposure, participants learn to expect a certain cue or pattern of cues to be observed, and act as if it had occurred when it was not present in the task.

The third type of error, that of using inappropriate heuristic strategies is one that can be found across previous experiments. Experiment 5 found evidence that some participants were using recognition heuristics which lead to failure on complex problems. In previous experiments, when establishing a mental set in a participant and that participant then uses enhanced cue perception to solve problems based on recognition heuristics, then this is one of the main reasons why performance is decreased on complex problems. However, if an inappropriate strategy is being employed, this can be moderated by the presence of feedback,

introduced with Experiment 2. It may be that more of these heuristic strategies were being used more in Experiment 1 – which had the lowest solution rates found for complex problems across all experiments, as feedback was not available. When feedback was introduced in Experiment 2, there was an overall increase in solution rates across all problem types compared to those seen in Experiment 1. It could be that when using these heuristic strategies, and the strategy fails, and participants receive feedback that the response was incorrect, that this then moves participant away from such strategies. The errors identified in Experiment 5 offer a descriptive narrative of the problem-solving approach used. While the qualitative methodology and low sample size may make Experiment 5 stand out from the other experiments, it greatly aids the conclusions of all other experiments by finding describable errors and accounts for the pattern of results seen across previous experiments.

8.2 Establishing the anomaly handling paradigm

One of the major aims of this thesis was to develop a novel paradigm for testing anomaly handling. The construction of such a task was detailed in Chapter 2, where the merits of using simulated environments for decision-making research was discussed. We claim the anomaly-handling paradigm presented throughout this thesis is robust and complex enough for use in decision-making research. The paradigm has strong evidential support to be an effective tool in investigating decisions. This is claimed because our expectations in relation to how participants would perform on the task were largely confirmed, with these expectations being derived from psychological literature and research on real world decision-making. That we could replicate the nature of mental set, and found success in describing our results using the RPD framework shows that behaviour on our task mimics real life decision-making.

The rules governing the anomaly handling paradigm were theorised to interact

creating both simple and complex scenarios. Throughout all of the experiments presented in this thesis, a significant main effect was found for complexity, where complex scenarios were solved less often than simple scenarios. This was a vital element of the development of a simulated environment. If all potential permutations of the systems were as easily solved as each other, then this would not have satisfied the requirement of being able to create anomalies in the form of complex problems.

The system developed for use in these experiments was able to successfully create these anomalous scenarios using extremely limited information. The system only has two streams of data available to participants; the weapons distributions, and the environmental conditions. Despite access to such basic information, the rules still interacted in such a way to allow complex problems to exist, rather than creating a simple simulation that offers no ability to further examine complex decision-making. A benefit of using a low information environment was the ability to use the paradigm on novice populations. All of the participants used in studies presented in this thesis were either university undergraduates, or wider members of the general population from around the world recruited online. No specific or specialist requirement was necessary to take part in these experiments. No computing or logic-based training was required to take part, and no specialist knowledge of intelligence, defence, security or conflict analysis was required to take part. As a result, the evidence shown throughout the described experiments shows that most participants were able to fully comprehend the nature of the simulated environment very quickly as evidenced through high solution rates for simple problems throughout all the experiments.

Another successful factor in the design and development of this anomaly- handling paradigm is its flexible nature. The version of the task used in this thesis relies upon a system of supply and demand – the logical rules for this were described in Chapter 2. Fundamentally, a finite resource is shared between a known number of locations, where the known local

conditions create predictable abnormalities in supply of the resource that lead to a possible determination between two causal events. The logic that underlies this system can be intact while changing a number of variables concerning the system itself. For example, the version of the task used in the described experiments used one resource (weapons) that were being supplied to four locations. Each change in supply due to the predictable abnormalities affected each of the locations equally.

The system is flexible enough to allow a wide variety of changes while still keeping the functional core the same to allow analysis of decision-making skills in the anomaly-handling environment. For example, the number of locations could potentially be increased to any desired level to represent a large area, the spatial relationship between the locations and the environmental blocker terrain features can be easily adjusted to represent actual geographic areas, and the specific post-event redistribution routes can be changed to represent local political alliances. All of these possibilities are capable of being implemented and represented within the simulated environment created for testing within this thesis. The system state shown in these experiments was programmed to be simple to allow the use of novice participants. The system was designed to be entirely modular, flexible and expandable so that any future research could be tailored to fit the needs to the participant population being tested. With the correct implementation, this anomaly-handling paradigm would be appropriate for the testing of experts in the security and defence industry providing more valuable insight into real world anomaly handling performance.

This is an important distinction because while we have replicated known psychological issues with decision-making throughout this research, it was found while using naïve populations in laboratory settings. The failure to successfully solve our anomaly problems, while analogous to the false-positive of drone-striking an innocent target, does not match the training, cognitive ergonomic environment, or the safety procedures set-up within a

military environment. While we have shown that it is still possible to commit cognitive errors in the identification of anomalies, and we know there are instances within the military where rare events were failed to be interpreted correctly, it is difficult to determine what contribution our anomaly-handling findings had on the same tragedies we aimed to better understand. Drone strike errors, or any industrial anomaly handling error, may not be the result of a cognitive misclassification due to reliance on similar cue characters, but may also result from bad intelligence, a lack of attention, or a number of other reasons. So while we have been successful in matching some theoretical errors with performance on a simulated task, it is not possible to claim that these are the only reasons for failure in the real world.

8.3 Boundaries of successful anomaly handling

The second major aim of this thesis was to investigate the circumstances under which successful anomaly handling could occur. This was measured by the ability of our participants to successfully solve the complex problem types presented to them within our anomaly handling paradigm.

Experiment 1 has shown that the ability to successfully interpret anomalies was diminished after exposure to a mental set. In this instance, the mental set was a concurrent series of problems that shares a single prominent familiar feature with that the complex problem. It was theorised that after exposure to this mental set, participants developed simple heuristics they used to aid in classification of the problem types, and that these heuristics centred on the same familiar feature present in complex problems. When participants subsequently apply the use of these heuristics in the incorrect classification of complex problems, they do so because they do not realise that the complex problems differ from the problems experienced previously during the establishment of the set.

However, unlike previous set-making research, where sets were established using similarity in solution-approach (Luchins & Luchins, 1959), familiarity with problem-space

(Bilalic, McLeod & Gobet, 2008), and knowledge structure (Wiley, 1998) the sets in our anomaly handling paradigm were established using perceptual similarities that existed between simple problems and complex problems. We suggest that the establishment of a set is due to participants over-attributing an importance to a single cue within the information environment and fail on complex problems when this cue is no longer diagnostic of the correct solution. This supports the previous research of Klayman (1988) who found that participants when given multiple cues, will attribute their own weighting of importance to a cue based on the likelihood of an outcome. As the majority of our problems were simple, and a single cue was sufficient, participants relied on this to solve complex problems when that cue alone was no longer sufficient.

We believe the results of our anomaly handling research can be successfully explained by the application of the RPDM framework. The RPDM framework suggests that problem-solvers will first compare a problem to similar problems experienced in the past. In this case, complex versions of an event appear more similar to the simple problems of the opposite event (i.e. complex conflicts appear similar to simple environmental events, and complex environmental events appear similar to simple conflicts). This similarity is based on the dependency of the cues determined above. Therefore, when participants experience a complex event, the most similar event available to that participant will be one that is of the wrong type. If solutions given by participants are based on their prior experience with the simple problems, then this will result in an error. This may be why we have seen lower solution rates for complex problems. Additionally, the RPDM framework may assist in the explanation of the effect of mental set. When a participant in the set condition is repeatedly shown a single problem type, they are better able to retrieve this from their repertoire of experience during the onset of the complex problem thus interfering more with the ability to generate a correct solution.

When comparing these findings to Experiment 2, the RPDM framework (Klein, Orasanu, Calderwood & Zsombok, 1993) is useful for explaining the pattern of results that were detected. It was found in Experiment 2 that introducing feedback was able to eliminate the effect of mental set. If, as RPDM suggests, participants compared new problems with old problems, and they are failing because they are comparing different types of event. When the feedback was shown to a participant after a wrong answer, they would find out that the problems in fact were not similar which prevented them from using this as a strategy again, whereas no such inoculation was provided in Experiment 1.

However, while eliminating mental set, solution rates for complex problem were still lower than simple problems, still demonstrating a lesser ability to deal with the anomalous events of the task. However, solution rates were increased in Experiment 2 compared to Experiment 1 and this again could be explained by the literature on cue acquisition and the RPDM framework. While errors are made, and feedback is given, participants will have used cues from the task to associate that trial with previous trials of a different type. As an error has occurred, participants will now understand this new case was not like the previous cases they had matched it to, and so the dependence of a single cue to make that attribution should be lower in subsequent presentations of that problem. As such, participants are building a separate repertoire of what complex problems look like, and learn that to develop this, they must pay attention to additional cues. This allows the explanation of why complex problems were solved at a higher rate in Experiment 2 (89%) than Experiment 1 (69%).

Experiment 3 attempted to find other ways of improving performance on complex problems. It was found that by introducing and highlighting a cost of failure, then performance improved across all problem types. This is an important finding, as anomaly handling often occurs in high-risk environments where the consequence of failure is

potentially huge. However, if decision-makers and anomaly handlers in these industries are aware of such consequences then this may provide some measure of inoculation against failure. This matches the research on accuracy/effort trade-off found by Bettman, Johnson, Luce & Payne (1993). If we are to be more accurate in a task, then that will take more effort. Conversely, if lower effort is put into a task, this will result in less accuracy. While this may sound straightforward, this provides an exciting opportunity to bring the naturalistic RPDM framework and the theoretical cue acquisition literature together again. If success on the task can be measured by the participant matching a problem with previous problems, and we have already established that errors are due to participants matching the wrong type of problem. Then we can suggest that the higher solution rates when there was a penalty to failure was due to participants better matching the new problem with previous problems. As the complex problems were designed in a way that such a match relied on multiple cues, then we can assume that the better performance when there was a penalty was due to participants using more cues. As participants who faced a penalty were more accurate in these assessments, it suggests that they were expending more effort – a finding also supported by their increased solution times. This means that when more effort was put into the task, participants were better able to pay attention to multiple cues. As attention was paid to these cues, it allowed their repertoire of experience to not make the attribution errors found in Experiment 1 and 2 at the same rates. This means that a heightened awareness to the cues, as a result of more effort put into the task via the threat of penalty, was able to develop a more sensitive model of recognition. While this is interesting for our research, it may also mean that the failures found on our task were often a result of participants not utilising the necessary effort to discriminate between cues. This may be good news to real-world anomaly-handlers such as drone operators or intelligence analysts who would be aware of the significance of their work, give it the attention and effort it requires, and thus offer defence against errors.

While the nature of a penalty improved performance, data from Experiment 3 also show a positive effect for mental set in the solution times for solving simple conflict problems. This type of problem was the same problem type used to establish the mental set. This shows that after exposure to a mental set, those participants are better able to identify problems that belong to the same family of problems as the mental set. It was theorised that this could be due to familiarity of what the set-making problems look like. For example, after experiencing a series of continuous simple conflict problems, participants develop awareness for what a conflict looks like, their repertoire of conflict is well established. This knowledge can then be later used to identify conflict problems faster than those participants who do not develop such extensive familiarity with a single problem type.

Experiment 3 also found a benefit for mental set in the solution rates for complex problems when there was no cost. As the no cost group was believed to be using suboptimal mental heuristics as discussed previously, it could be that this familiarity with a single problem type confers benefits not only to faster recognising problems of the same type, but in also helping to identify when a problem is not the same. By becoming familiar with what a conflict looks like, when encountering the complex environmental problems that have the same weapons distribution as a conflict, participants in the set group know from experience that this is not a conflict as it does not match prior examples. In Experiment 1, there were only two blocks of problems, and so participants did not have the time to develop this familiarity, and so mental set acted a negative, as participants used prior experience as a misplaced solution aid, categorising the anomaly problems based on superficial details and approximate likeliness. However, in Experiment 3, there were six problem blocks giving participants three times as many conflicts to become familiar with. In this experiment, it is likely that participants have more time to establish a familiarity with the set-making problem and so can use this to their advantage. It is important to note that this benefit of mental set

only occurs when potentially using sub-optimal heuristic strategies. When there was a penalty, it was believed that participants slowed down and took more care with their decisions, reflecting a more optimal approach due to more sensitivity to what type of problem they were encountering. For these cases where a penalty was present, there were no differences between set and no set groups.

This notion of familiarity conferring a benefit was tested further in Experiment 4. It was found that when the problem type used to establish the mental set was different from the most frequently occurring problem type, that those exposed to the set were better able to solve simple problems. This could be used to support the argument in Experiment 3 that exposure to a mental set helps develop a familiarity, which can further be used to aid correct classification.

However, solution rates for complex problems in Experiment 4 indicate that the more similar the set-making problem type is to the complex problem that comes after it, the less likely participants were to obtain a successful solution. This shows that while familiarity can be a helpful heuristic recognition tool used to classify the more distinct problem types, that mental set is still a burden when it comes to distinguishing perceptually similar but functionally different environments.

Previously, mental set has only ever been explored within the psychological literature as a barrier to successful problem solving. Mental rigidity is a state that is universally discussed as conferring a negative to those that possess it. However, we have found a novel and adaptive use for such a phenomenon. We have found that mental set confers the benefit of establishing when something looks similar to something previously seen. This can have a number of benefits, such as the saving of cognitive effort when utilising adaptive heuristics such as the implementation of the data-frame model, and the RPD framework that suggest critical decision by industrial experts are often made by comparing their observations against

a repertoire of prior experience. Perhaps this mental set is a faster way of developing such an experience, and that it may be beneficial in the majority of cases when natural events are distinct from each other, it may only lead to negative consequences when there are high levels of similarity between observations and experience. While it is novel to successfully explore a positive benefit of mental set in relation to successful anomaly handling, it appears that the more similar an anomaly is to its typical environment, the harder it may be to spot.

8.4 Applications

In terms of real world application for anomaly handlers and high risk decision makers, a number of possible applications and possible recommendations can be made on the basis of the studies described in this thesis.

This research aimed to develop an anomaly handling simulation as a dynamic training tool for professions where pattern recognition and interpretation is required, and may be used to expedite the route to expertise in an environment where mistakes can be made safely. As discussed previously, the simulation developed for use in this thesis is a robust tool for creating anomaly problems. The context, domain, and difficulty of the simulation are all possible to be adjusted while still keeping the logical system rules in place that allow decision-making to occur. This system could be altered to meet specific situational training, could be delivered as an anomaly handling tool to teach more complex decision-making processes in general, and also demonstrate the negative effects of those things which impact upon successful decision making (mental set, unfamiliarity with environment, lack of awareness of cost).

There is also the potential to develop decision-support tools that may increase the performance of professionals, where the consequences of failure are potentially catastrophic. In analytical fields systems could be developed that warn system users when they have been operating a similar environment for too long, as this could foster the development of a mental

set. Or additional process analysis could be done to integrate new business requirements such as alarms, when a small variance is detected from an otherwise similar state, as this is when error is more likely to occur.

8.5 Future research

Based on the findings in these studies, a number of potential future areas for research are worth mentioning. Firstly, it would be important to the future success of this anomaly handling paradigm for it to be tested by experienced and expert decision makers from the same industry as the simulation is contextualised with. It would provide greater evidence for the value of the paradigm if expertise can be transferred from real world decision making tasks to success on the problems delivered using our simulation. This would establish the validity of the paradigm when discussing real world decision makers and make any conclusions about strategy use more applicable to the wide industrial audience. Testing experts would also allow new opportunities for the paradigm to be reconfigured into a larger and more complex simulated environment.

Secondly, a number of theoretical conjectures were made in the analysis of results across experiments that would require resolution. While mental set provided a negative effect in Experiment 1 (due to similarity between set problems and complex problems), it provided a positive effect in Experiment 3 and 4 (due to stimuli familiarity). While the strongest reasons for these were theorised above, it would be beneficial to further examine the nature of these differences. This could be achieved using verbal protocols where participants talk through their decision making process. This would allow an evidence based rationale for why errors are being committed on complex problems and could test the current hypotheses stated for failure in this thesis.

Thirdly, the paradigm used in this thesis aimed to be representative of systems used by anomaly handlers and decision makers in the real world. It would be to the value of this

research if more real world research could be undertaken that shows how decision makers operate in their natural environments to ensure that the paradigm is keeping pace with the industries in which it attempts to replicate. The findings found using this paradigm should also be validated in more contextualised scenarios which would require the study of real world environments and the decisions that are being made within them. A series of ethnographic studies should be completed that follow expert decision makers into their respective industries and attempt to validate the conclusions drawn from research using this paradigm. This would allow us to comment on the real extent that stimuli familiarity, or saliency of cost, or susceptibility to mental set, would have on real world decision makers for whom this research was intended to benefit.

8.6 Conclusion

In summary, this thesis has presented a series of decision making studies using a novel paradigm to test a novel concept of anomaly handling. It has drawn on previously utilised psychological research, including mental set, as an attempt at understanding the failure of successful decision making for anomalous problems.

It has been shown that the paradigm was successfully tested in a number of different iterations, from under laboratory-controlled settings using undergraduate psychology students, to online implementations that reached a global audience. This research has been a successful example of the development of a simulation for use in psychological research, and has offered a number of insightful findings around the nature of strategy use in such a complex environments, and has important consequences for those industries in which failure to handle anomalies correctly is not an option.

9 References

- Ackerman, S. (2016, July 1). US to continue 'signature strikes' on people suspected of terror links. *The Guardian*. Retrieved from www.theguardian.com on 03/12/2017.
- Ball, L., & Ormerod, T., & Morley, N. (2004). Spontaneous analogising in engineering design: A comparative analysis of experts and novices, *Design Studies*, 25(5), 495-508.
- Bettman, J., Johnson, E., Luce, M., & Payne, J. (1993). Correlation, Conflict, Choice, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19 (4), 931-951.
- Bilalić, M., McLeod, P., & Gobet, F. (2008). Inflexibility of experts—Reality or myth? Quantifying the Einstellung effect in chess masters, *Cognitive Psychology*, 56(2), 73-102.
- Briggs, G., Hole, G., & Land, M. (2011). Emotionally involving telephone conversations lead to driver error and visual tunnelling, *Transportation Research, Part F (14)*, 313-323.
- Cowan, D. (1986). Developing a process model of problem-recognition, *Academy of Management Review*, 11 (4), 763-776.
- Crandall, B., & Getchell-Reiter, K. (1993). Critical decision method: A technique for eliciting concrete assessment indicators from the intuition of NICU nurses. *Advances in Nursing Science*, 16(1), 42-51.
- Cullen, W. (1990). The public inquiry into the Piper Alpha disaster, vols I and II. HMSO, London.
- Dawes, R., & Corrigan, B. (1974). Linear models in decision making, *Psychological Bulletin*, 81(2), 95-106.
- Dawes, R., Faust, D., & Meehl, P. (1989). Clinical versus actuarial judgment. *Science*, 243 (4899), 1668-1674.

- De Keyser, V., & Woods, D. (1993). Fixation errors: Failures to revise situation assessment in dynamic and risky systems. In Colombo, A., & de Bustamente (Eds.) *Advanced systems in reliability modelling*. Kluwer; Norwell.
- Englehardt, T. (2013, December 20). The US has bombed at least eight wedding parties since 2001. *The Nation*. Retrieved from www.thenation.com on 03/12/2017.
- Ericsonson, K., & Simon, H. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, 5(3), 178-186.
- Evans, J. (1989). *Bias in Human Reasoning: Causes and Consequences*. Brighton; Erlbaum.
- Fasolo, B., McClelland, G., & Lange, A. (2005). *The effect of site design and interattribute correlations on interactive web-based decisions*, in C. Haugtvedt, K. Machleit and R. Yalch (Eds.) *Online Consumer Psychology: Understanding and Influencing Behavior in the Virtual World*, 325–342, New Jersey: Erlbaum.
- Fiske, S., & Taylor, S. (1991). *Social Cognition*. New York: McGraw-Hill.
- Friedman, D., & Massaro, D. (1998). Understanding variability in binary and continuous choice. *Psychonomic Bulletin & Review*, 5(3), 370-389.
- Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making, *Annual Review of Psychology*, 62, 451-482.
- Gonzalez, C., Vanyukov, P., & Martin, M. (2005). The use of microworlds to study dynamic decision making, *Computers in Human Behaviour*, 21(2), 273-286.
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, 12(1), 19-30.
- Grove, W., & Meehl, P. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy, *Psychology, Public Policy, and Law*, 2, 293-323.

- Hilburn, B., Jorna, P., Byrne, E., & Parasuraman, R. (1997). The effect of adaptive air traffic control (ATC) decision aiding on controller mental workload, *Human-automation interaction: Research & Practice*, 84-91.
- Hoffman, R., Crandall, B., & Shadbolt, N. (1998). Use of the critical decision method to elicit expert knowledge: A case study in cognitive task analysis methodology, *Human Factors*, 40 (2), 254-276.
- Johnson, R., Stone, B., Miranda, C., Vila, B., James, L., James, S., Rubio, R., & Berka, C. (2014). Identifying psychophysiological indices of expert v novice performance in deadly force judgement and decision making, *Frontiers in Human Neuroscience*, 8, 512.
- Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioural economics, *American Economic Review*, 93(5), 1449-1475.
- Kahneman, D., & Tversky, A. (1973a). Availability: A heuristic for judging frequency and probability, *Cognitive Psychology*, 5(2), 207-232.
- Kahneman, D., & Tversky, A. (1973b). On the psychology of prediction, *Psychological Review*, 80(4), 237-251.
- Kahneman, D. & Tversky, A. (1972). Subjective probability: a judgement of representativeness. *Cognitive Psychology*, 3, 430-452.
- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, 49, 81.
- Klayman, J. (1988). Cue discovery in probabilistic environments: Uncertainty and experimentation, *Journal of Experimental Psychology; Learning, Memory, and Cognition*, 14(2), 317-330.
- Klein, G.A. (1989). Recognition-primed decisions. In W. Rouse (Ed.), *Advances in Man-*

- Machine Systems Research*, 47-92. JAI Press; Greenwich, CT.
- Klein, G. A., Calderwood, R., McGregor, D. (1989). Critical decision method for eliciting knowledge. *IEEE Transaction on System, Man and Cybernetics*, 9 (3).
- Klein, G., Orasanu, J., Calderwood, R., & Zsombok, C.E. (1993). *Decision making in Action: Models and Methods*. Ablex. Norwood, NJ.
- Klein, G., Pliske, R., Crandall, B., & Woods, D. (2005). Problem detection, *Cognition, Technology and Work*, 7 (1), 14-28.
- Klein, G., Moon, B., & Hoffman, R. (2006). Making sense of sensemaking 2: A macrocognitive model, *IEEE Intelligent Systems Conference Proceedings*, 21(5).
- Klein, G., Philips, J., Rall, E., & Peluso, D. (2007). A data/frame theory of sensemaking. In R. Hoffman (Ed.) *Expertise Out of Context, Proceedings of the 6th International Conference on Naturalistic Decision Making*. Erlbaum; NJ.
- Lagnado, D., & Slowman, S. (2004). The advantage of timely intervention, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(4), 856-876.
- Luchins, A. S. (1942). Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs*, 54(6).
- Luchins, A. & Luchins, E. (1959). *Rigidity of behaviour: A variational approach to the effect of Einstellung*, Oxford, UK; Oregon Press.
- Malakis, S., & Kontogiannis, T. (2012). A sensemaking perspective on framing the mental picture of air traffic controllers, *Applied Ergonomics*, 44(2), 327-339.
- Mednick, S. (1962). The associative basis of the creative process, *Psychological review*, 69(3), 220.
- Miller, S., Appleby, S., Garibaldi, J., Aickelin, U. (2013). Towards a more systematic approach to secure systems design and analysis, *International Journal of Secure Software Engineering*, 4(1), 11-30.

- Morley, N., Ball, L., & Ormerod, T. (2006). How the detection of insurance fraud succeeds and fails, *Psychology, Crime & Law*, 12(2), 163-180.
- Mumaw, R., Roth, E., Vincente, K. & Burns, C. (2000). There is more to monitoring a nuclear power plant than meets the eye, *Journal of Human Factors and Ergonomics Society*, 42(1), 36-55.
- Newell, B., Lagnado, D., & Shanks, D. (2007). *Straight choices: The psychology of decision making*. Psychology Press.
- Norman, D. A. (1991). Approaches to the study of intelligence. *Artificial Intelligence*, 47(1), 327-346.
- Omodei, M., & Wearing, A. (1995). The Fire Chief microworld generating program: An illustration of computer-simulated microworlds as an experimental paradigm for studying complex decision-making behaviour, *Behavior Research Methods, Instruments & Computers*, 27(3), 303-316.
- Payne, J. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis, *Organizational Behavior and Human Performance*, 16(2), 366-387.
- Phelps, R.H., & Shanteau, J. (1978). Livestock judges: How much information can an expert use? *Organizational Behavior and Human Performance*, 21(2), 209-219.
- Reason, J. (1990). *Human Error*. UK; Cambridge University Press.
- Riddle, K. (2010). Always on my mind: Exploring how frequent, recent, and vivid television portrayals are used in the formation of social reality judgements, *Media Psychology*, 13(2), 155-179.
- Rogovin M. (1980). Three Mile Island: A report to the Commissioners and to the Public, *Nuclear Regulatory Commission, Special Inquiry Group*.
- Sarbin, T. (1943). A contribution to the study of actuarial and individual methods of

- prediction, *American Journal of Sociology*, 48, 593-602.
- Shanteau, J. (1992). How much information does an expert use? Is it relevant? *Acta Psychologica*, 81, 75-86.
- Smith, G. F. (1989). Defining managerial problems: a framework for prescriptive theorizing. *Management Sciences*, 35(8):963–981.
- Sneddon, A., Mearns, K., & Flin, R. (2006). Situation awareness and safety in offshore drill crews, *Cognition, Technology and Work*, 8, 255-267.
- Sturke, J. (2008, July 11). US air strike wiped out Afghan wedding, inquiry finds. *The Guardian*. Retrieved from www.theguardian.com on 03/12/2017.
- Stanovich, K., & West, R. (2000). Individual difference in reasoning: Implications for the rationality debate, *Behavioural and Brain Sciences*, 23, 645-726.
- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers, *Psychological Bulletin*, 76(2), 105-110.
- Tversky, A. & Kahneman, D. (1973). Availability: A Heuristic for Judging Frequency and Probability. *Cognitive Psychology*, 5 (2), 677-695.
- Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases, *Science*, 185, 1124-1131.
- Wafa, A.W., & McDonald, M. (2008, November 5). Deadly US airstrike said to hit Afghan wedding party. *New York Times*. Retrieved from www.nytimes.com on 03/12/2017.
- Wason, P. (1960). On the failure to eliminate hypotheses in a conceptual task, *Quarterly Journal of Experimental Psychology*, 12(3), 129-140.
- Wason, P. (1966). Reasoning. In Foss, M. (Ed.) *New Horizons in Psychology*. Harmondsworth: Penguin.
- Wason, P., & Johnson-Laird, P. (1972). *Psychology of reasoning: Structure and content*, Cambridge, MA: Harvard University Press.

Wason, P., & Evans, J. (1975). Dual processes in reasoning? *Cognition*, 3(2), 141-154.

Wiley, J. (1998). Expertise as mental set: The effects of domain knowledge on creative problem solving, *Memory & Cognition*, 26(4), 716-730.

Appendix 1. Screenshots of the online task instructions

Welcome!

In this task you will be asked to play a game where you must track the movement of weapons being smuggled into various zones of a fictional country.

You will be asked to monitor the number of weapons in each zone. There are certain events that can cause the number of weapons in each zone to change in specific ways.

These events are **conflicts** or **environmental events**. In some cases, it may be that no event is present.

Your task is to identify which event, if any, is causing the observed change in the number of weapons in each zone, and to assess if a conflict is occurring.

By pressing the button below you will begin training on how to detect each type of event, and will also introduce you to the interface used throughout the experiment.

Begin Training

© 2023 www.anomalyhandling.com

Instructions

Reading the Map

In this task, you will play the role of an intelligence analyst.

You are tasked with monitoring a fictional unstable country. A map of this country can be found to the right. This map will be present during all of the scenarios so you do not need to memorise it.

The country has **four major population zones marked A, B, C and D**.

The **red lines represent the supply routes** used to traffic arms into each of these zones.


In this task, your goal is to **examine the movement of weapons** around the country, and to use this information to **detect conflicts**. This is done by spotting particular patterns in arms traffic, you will receive training on how to detect conflicts shortly.


It is sometimes difficult to spot a conflict due to the volatile nature of the environment. The environmental conditions can affect the supply of weapons, and you will receive more training on this shortly


Also on the map are various **terrain features** that must be crossed by some supply routes.

To help you remember this, there are symbols on the map that show which terrain features must be crossed to reach each zone.

Key:

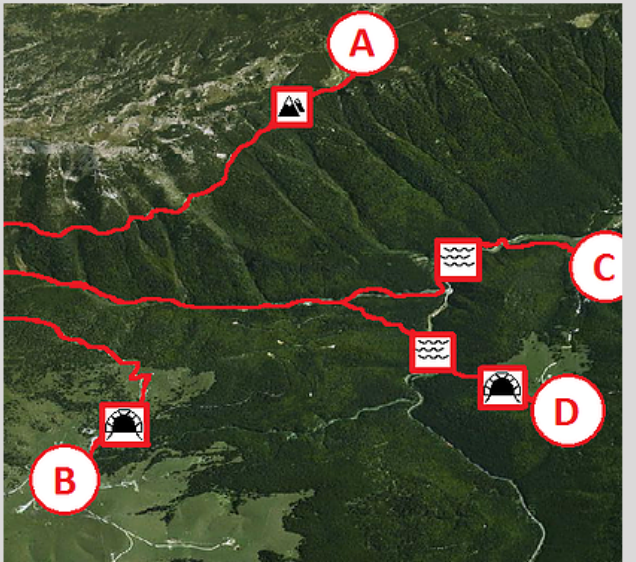
Routes that cross the mountain are shown using the  symbol.

Routes that cross the river are shown using the  symbol.

Routes that use the underground tunnels are shown using the  symbol.

When you are familiar with how to identify the components of the map, please click the 'Next' button to move onto the next set of instructions.

Next



Instructions

Identifying Conflicts

During the task, you will have access to other information to help you make your decision.

One type of information is the **number of arms being trafficked into each zone**.

This information is contained in a table like the ones to the right. It is possible to see the number of arms being supplied to each zone across time.

Due to the unstable nature of this environment, it is not unusual for arms to be trafficked to all zones.

The top table on your right shows what a 'normal' day looks like. This is the typical amount of arms being trafficked on any given day.

During a 'normal' period, arms are distributed equally between all zones. By looking for changes in this distribution, it is possible to identify when a zone is in conflict.

When a zone enters a state of conflict, the demand for weapons in that zone increase. This causes half of the supply from other zones to be redirected towards the conflict zone.

An example of what the weapons distribution would look like during a conflict can be seen in the bottom table on the right. Although the weapons are distributed equally at the start, there is an accumulation of weapons in Zone B over time that would indicate a conflict is occurring in that zone.

During the task, **all arms traffic will begin being spread equally and may change over time**. By spotting this change you can successfully identify when and where a conflict is occurring.

It is important to note that there are only ever enough weapons to provide for one conflict at a time.

Conflict Indicator - Estimated Weapons

Zone A	Zone B	Zone C	Zone D
403	399	393	404
392	393	398	400
402	403	404	404
405	408	406	406
407	395	405	398

Start



Time



Finish

Conflict Indicator - Estimated Weapons

Zone A	Zone B	Zone C	Zone D
399	392	396	404
403	395	395	402
328	604	327	328
247	789	252	246
203	1016	204	204

Start



Time



Finish

Once you are familiar with how to read the table, and understand how to identify a conflict, please click the 'Next' button. Click the 'Back' button to return to 'Reading the Map'.

Back

Next




Instructions

Environmental factors

The last page showed you how to read the number of weapons that are supplied to each zone, and how to identify a conflict.

The last remaining information you will be given is that of the environmental conditions within the country. You will be given a table of three environmental factors - rainfall, temperature, and seismic activity. An example of these conditions are displayed in a table on the right.

Each of these factors interact with the terrain features from the map in a unique way:

- **High rainfall** will flood the river 
- **Low temperatures** will freeze the mountain 
- An **increase in seismic activity** will collapse the underground tunnels 

When environmental factors affect a terrain feature, they become treacherous to use. Supply routes that cross a treacherous terrain feature are only able to deliver half of their weapons.

When environmental factors prevent the supply of weapons, all undelivered weapons will be redistributed along all remaining safe supply routes.

There will be examples of how this works on the next page.

You will not be told at what level the environmental conditions become treacherous, but it is possible to discover by the change over time, and the effect on weapon supply.

An example of how treacherous environmental conditions can develop over time can be seen in the bottom table on

Local Environmental Conditions

Rainfall (mm)	Temperature (°C)	Seismic activity
1	17	2.9
4	10	2.3
9	17	2.3
7	11	2.8
8	13	2.2

Start



Time



Finish

Local Environmental Conditions

Rainfall (mm)	Temperature (°C)	Seismic activity
5	18	2.7
11	0	3.8
23	-4	4.9
45	-11	5.5
67	-23	6.7

Start



Time



Finish

You should now be able to get information on the environmental conditions, and which environmental factors affect each terrain feature.

To see examples of how environmental factors affect weapon supply, please click 'Next'. Click 'Back' to return to 'Identifying Conflicts'.

Back

Next

Summary

Map

- There are four zones marked A, B, C, and D.
- Supply routes to each zone are marked.
- The symbols on the map indicate which terrain features a supply route must cross:



Mountain



River



Underground tunnels

Weapons

- Weapons start off equally distributed between all zones.
- Over time, weapons distribution may change.
- During a conflict, weapons will accumulate in one zone, reducing all non-combat zones supply by half.
- Only one zone can be in conflict at the same time.
- Zones using treacherous supply routes only receive half weapons supply, with remaining supply distributed along all safe routes.
- The number of weapons will change between timepoints, small changes are normal. Conflicts can be detected by spotting large changes.

Environmental conditions

- Environmental conditions affect terrain features:
 - **High rainfall** floods the river
 - **Low temperature** freezes mountain
 - **Increase in seismic activity** collapses underground tunnels
- Supply routes become treacherous if they must cross an effected terrain feature.
- The environmental conditions will change between timepoints, small changes are normal. Treacherous conditions can be spotted by drastic and rapid changes.

Back

Next

Appendix 2. Trial order for Experiment 4.

Trial order over 3 blocks for No set (N), Simple set (S), and Complex set (C) conditions.

Legend

